Presidential Address

Psychological Assessment Versus Psychological Testing

Validation From Binet to the School, Clinic, and Courtroom

Joseph D. Matarazzo *Oregon Health Sciences University*

ABSTRACT: Increasingly, psychological assessment is conducted with clients and patients involved in child custody and personal injury litigation. Clinical neuropsychologists are being asked sophisticated questions by attorneys regarding the validity of practitioners' most highly respected tests. Research reviewed here bears on the validity of test-buttressed clinical opinions, including research related to the following psychometric properties of individual test scores: standard errors of measurement, test–retest stability and subtest-to-subtest intercorrelations. The highest and the lowest subtest scores used as indices, respectively, of an individual's premorbid level of cognitive functioning and the degree of current impairment from that presumed earlier level is not justified when used in isolation from the life history and current medical findings. Although many practitioners use information from the wider research, courtroom experience suggests that a number do not; contrariwise, the attempt of Faust and Ziskin (1988a) to undermine the courtroom testimony of every psychologist who serves as an expert witness is also criticized.

Historical Roots of Psychological Assessment

Authors of textbooks in psychology typically date the beginning of psychological assessment with the works of Francis Galton, James McKeen Cattell, Lightner Witmer, Alfred Binet, and other psychologists who published their works in the last decades of the 19th century. However, Doyle (1974) has quoted passages from Plato that indicate that individual assessment for the purpose of selecting young men for state service on the basis of individual differences in both mental abilities and physical abilities was practiced in Ancient Greece 2,500 years ago.

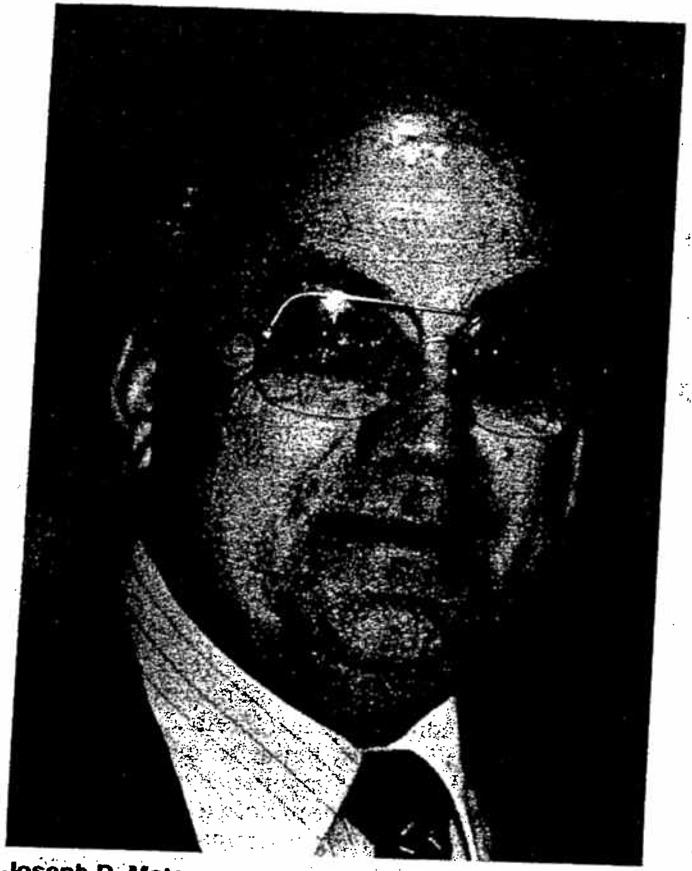
A parallel but apparently independently developed system of assessment and selection also existed in Ancient China. Specifically, DuBois (1970) and Bowman (1989) cited historical documents indicating that circa 200–100 B.C.—2,000 years before Alfred Binet and Theodore Simon constructed the first modern tools used in psychological examinations of children and adults today—remarkably effective objective tests were being used by local authorities under the direction of the Emperor of China. These tests were used to select the most able applicants for civil service positions in municipal, county, provincial, and national government. These earliest tests measured various aptitudes, including level of literacy, verbal cleverness, writing, arithmetic, civil law, revenue, and geography of the Empire. In a recent commentary on that era of mental testing in Ancient China, Bowman wrote

> The changes in the examinations (from Ancient China to modern times) across hundreds of years of experience, controversy, and reform reveal interesting testing history . . . including issues that foreshadow some of the controversies in ability testing in modern times. Such topics as the relative importance of memory as a feature of mental ability, the role of expert knowledge, effects of social class on test performance, the use of examinations to provide opportunities for social mobility, personal recommendations as an alternative to formal examinations in personnel selection, social protest against the nature of the examinations, the use of geographical units in allocating quotas of candidates to be passed, and the need to measure applied problem solving and reasoning were all vigorously debated. (p. 578)

Although achievement tests used for selection in Ancient China and Greece are not comparable to today's tests of global intelligence ushered in by Binet and Simon, aptitudes tested in China two millenia ago are known to correlate highly with aptitudes measured by modern tests of intellectual and cognitive ability (J. D. Matarazzo, 1972, pp. 45–47, pp. 245–247).

Thus, it is of more than passing interest that academic and industrial psychologists and those who used tests for selection in schools and industry, and others associated with our country's testing industry during the past 80 years, continue to be embroiled in the same types of controversies that befell generations of test users dating back to Ancient China. Modern intelligence tests, such as Terman's individually administered 1916 Stanford Revision of the Binet-Simon and the current offshoots of the group-administered Otis test that after 1919 (J. D. Matarazzo, 1972) were used in our public schools for assigning children to curricula, or tracks, on the basis of test-diagnosed ability groupings, also underwent vehe-

Copyright 1990 by the American Psychological Association, Inc. 0003-066X/90/$00.75
Vol. 45, No. 9, 999–1017



Joseph D. Matarazzo

ment attacks not unlike those leveled at the tests used by the Chinese. Initially, the modern-day attacks came from a political commentator, Walter Lippmann (1922), and a handful of psychologists of his era who believed in a more egalitarian system of public education (Sokal, 1987) than the tracking system then being instituted on the basis of test scores. During the past 25 years the attacks have taken the form of laws and administrative regulations enacted by the U.S. Congress and some state legislatures to protect the rights of public school children and of adults in the work force, who were believed to be subject to injury from the improper use of psychological testing (for excellent reviews see Amrine, 1965; Elliott, 1987; Cohen, Montague, Nathanson, & Swerdlik, 1988).

Psychological Testing Versus Psychological Assessment

Until the last decade, our country's clinical psychologists were spared the public scrutiny, criticism, and statutory regulation that befell their school and industrial-psychologist counterparts. However, recent changes in the laws and regulations—and especially in the practices of attorneys associated with child custody and personal injury litigation—have caused a shift in the perception of the clinical psychologist as an ally and provider of assessment services that would benefit the examinee to one in which the psychologist is a potential adversary. That is, a clinician is hired by an attorney who represents only one of the parties in the litigation and who hopes that the results of the psychological examination will support the interests of his or her client (i.e., a husband and not the wife, or an insurance company and not the examinee, or vice versa).

As a consequence, an increasing number of attorneys recognize that even in our nation's most advanced centers for psychological assessment, the measurement of intelligence (or personality, memory, or other psychological functions) is not, even today, a totally objective, completely science-based activity. Rather, in common with much of medical diagnosis, experience in our nation's courtrooms is forcefully making clear to psychologists that the assessment of intelligence, personality, or type or level of impairment is a highly complex operation that involves extracting diagnostic meaning from an individual's personal history and objectively recorded test scores. Rather than being totally objective, assessment involves a subjective component. Specifically, it is the activity of a licensed professional, an artisan familiar with the accumulated findings of his or her young science, who in each instance uses tests, techniques, and a strategy that, whereas also identifying possible deficits, maximizes the chances of discovering each client's full ability and true potential.

Competent practitioners in psychology learn from clinician role models during apprenticeship training and from their own subsequent experiences that, objective psychological *testing* and clinically sanctioned and licensed psychological *assessment* are vastly different, even though assessment usually includes testing. Personnel technicians, elementary school teachers, and high school counselors monitoring, respectively, a group administration of the Otis Classification Test, Iowa Tests of Educational Development, or College Entrance Examination Board's Scholastic Aptitude Tests (SAT), and other tests described elsewhere (J. D. Matarazzo, 1972), are involved in psychological *testing*, an activity that has little or no continuing relationship or legally defined responsibility between examinee and examiner. Psychological *assessment*, however, is engaged in by a clinician and a patient in a one-to-one relationship and has statutorily defined or implied professional responsibilities. With the exception of those examiners involved in litigation, the typical psychological examination carried out by the clinical psychologist is geared specifically to the benefit and needs of the particular patient, determined from a careful reading of the patient's hospital chart or, in the case of an outpatient, from a telephone call or letter of referral.

This article is based on the presidential address delivered at the meeting of the American Psychological Association, Boston, MA, August 11, 1990. The article integrates and expands ideas and material previously published in disparate sources.

Correspondence concerning this article should be addressed to Joseph D. Matarazzo, School of Medicine, Oregon Health Sciences University, Portland, OR 97201-3098.

Assessing Deficit Versus Potential

Until recently psychologists in clinical settings strived to assess potential not deficit. Often the activity carried on outside a clinical setting, testing intelligence and other human attributes as one component of a selection decision, continues to be carried out to benefit other persons (e.g., college admissions personnel, employers) who, because of an oversupply of applicants, are searching for deficits and frailty among individual examinees in the hope of weeding out the ones believed to be less qualified. For the practicing clinical psychologist, however, whose statutorily defined focus of interest and professional responsibility is the individual examinee, the challenge in the assessment enterprise usually is to assess human potential, with interest in but relatively less emphasis on deficit. Thus, in a clinical setting each patient or client, with the exception of some court-mandated referrals, is provided assessment and other services within a framework of individual-oriented appraisal, with rehabilitation or an improvement in the human condition as the end.

Nevertheless, as some of our psychologist colleague-critics and federal and state legislators and judges have recently made clear, psychological assessment techniques, in common with most tools, can be used for many purposes, some harmful and some helpful, and their use cannot be separated from their validity and from the training, competence, and ethical values of the psychologist using them. In the hands of a good clinician, the results of an examination of intelligence or personality, correlated with information from the person's history, are as useful as analogous information would be in the hands of a good surgeon, internist, accountant, or plumber. In the hands of a fool—whether psychologist, physician, physicist, elementary school teacher, college admissions officer, surgeon, or plumber—such data are tools for potential harm.

With the exception of research published in psychological journals, until the 1970s information about the reliability and validity of the psychological assessment tools used by psychologist–clinicians was shared primarily with the individual patient or client served and with colleagues working in hospitals and clinics, with whom we pooled the information gathered during our clinical work. When I began my career in 1952, there were no effective treatments for any mental illnesses; it mattered little whether the diagnosis we gave a patient was schizophrenia, manic depression, or another disorder, inasmuch as the treatment (institutionalization) was essentially the same. For that reason, in hospital and clinic settings 40 years ago, even when a mistake was made, relatively little additional harm was done to those mentally ill patients.

Establishing Validity: Shift From Journals to Congress and the Courts

Given the advances in diagnosis and treatment in today's new era of litigation, the validity of clinical work and the mistakes clinicians make are increasingly a matter of public as well as professional record. For the first time, many clinicians find that not only are they no longer professionally responsible to, but are in a clear adversarial relationship with, the clients whom they are being asked to examine either by the plaintiff's or the defendant's attorney. As a result of the extraordinary human and financial costs involved in such legal actions, psychologists who are involved in professionally impersonal litigation are undergoing extremely fierce and highly sophisticated examination and cross examination regarding the scientific integrity of the same clinical psychological tools whose validity rarely was questioned by other service providers and patients when clinical psychologists practiced primarily in hospitals and clinics. As a result, more and more clinicians have been forced to go back to study carefully, and in some instances to totally reexamine, our earlier beliefs stemming from the published professional and scientific literature on which so much of our day-to-day professional work depends.

Changes that began in the 1950s in the types of examinees administered the Minnesota Multiphasic Personality Inventory (MMPI) shifted the potential impact of the products of the professional contributions of psychologists from patients to job applicants. Although the MMPI was developed in the 1930s as a clinical instrument for assessing the individual hospitalized patient, by the 1950s it also was being responsibly used by well-qualified psychologists for assessing not only the individual outpatient but also executives in industry. Unfortunately, by the early 1960s the MMPI also began to be used in isolation or with very limited supervision by untrained personnel clerks for hire/don't-hire decisions involving employee applicants. As a result, in 1964, Senator Sam Erwin, Jr. and Representative Cornelius Gallagher introduced and helped pass federal legislation to outlaw such use of the MMPI in employee hiring by our government (Amrine, 1965). Since then, other court decisions have outlawed, without proof of prior validation, the use of group and individual tests of intelligence in classroom placement of youngsters in the school systems of Texas, California, and other states, as well as with prospective employees in industry (Amrine, 1965; Cohen et al., 1988; Elliott, 1987).

These congressional and judicial decisions had a clear message for psychology: Given the human costs involved, in the event a mistake was made, society now wanted firmer evidence of the validity of opinions offered by psychologists in job hiring and in the schools. Society had spoken out 25 years ago that turning down a job applicant or placing a minority child or a poor child in a special education class for slow learners entailed human costs that were too high to be based solely on the professional belief of the consulting psychologist (or technician surrogate) that the tests, which formed a core part of his or her assessment decision, had been adequately validated.

Although severe roadblocks have been imposed between 1960 and the present by statutory and executive and judicial opinions on the further use of psychological tests in industry and schools, scientific and professional psychology has weathered these societal constraints reasonably well. This was due in part to the availability to psychologists of other, less controversial means of assessing job applicants and school children (e.g., achievement and other tests that were less general and more school- or job-skill specific).

Psychologists' Testimony in the Courtroom

Unlike the situation involving a rejected applicant or parents who allege that their child was placed in an inferior classroom program, in which the short-term costs of a psychologist's error are primarily human hurt to one individual, in one of today's personal-injury-initiated psychological-assessment consultations large sums of money also are involved. It is not unusual for a clinical psychologist or a clinical neuropsychologist to examine, at the request of an attorney, insurance company, or other payer, a person who alleges a brain injury or a stress disorder and whose request for redress involves millions of dollars. In the more traditional office practice of child and family psychology, psychologists are no longer examining only the school child who appears to be a slow learner; the healthy child in a custody battle, as well as each parent, have also become the focus of examination.

Given this new dimension that involves healthy children or huge sums of money, the legal profession is engaging increasing numbers of psychologist–clinicians in a debate being carried out in the courtroom and is forcing us to demonstrate without equivocation ("within reasonable psychological probability") the validity of our clinical opinions, including opinions based on our most respected instruments for psychological assessment. This recent experience has been a humbling one for psychologists. The newly involved attorneys, juries, and judges are asking psychologists in the courtroom for considerably more evidence than our clinic or hospital colleagues have requested to demonstrate that the instruments and techniques used, in part, in forming their clinical opinions are valid ones.

Psychologists and Patients as Adversaries

In place of referrals for the pre-1980 type of patient-oriented psychological diagnosis, attorneys, courts, and state workers' compensation and related agencies are requesting that psychologists assess a patient–client during one or two sessions and render an opinion in writing, with no continuing professional responsibility either to that patient–client or to the third-party payer or plaintiff who potentially pays or receives large sums of money.

In the weeks or months after a written opinion is rendered, an opinion that perforce will please one party to the litigation and psychologically devastate as well as materially harm the second party, the clinician–psychologist who made the assessment is forcefully confronted by a seies of actions triggered by the human and financial costs of his or her opinions. Typically, this begins with a court-reporter-recorded deposition, taken under oath, initiated by the attorney for the side that has been hurt by the opinion. Such a deposition permits the injured side to probe for strengths and weaknesses in the psychologist's rendered opinions. If the potential financial sums or human costs are substantial, in the interval between the deposition and actual jury trial or hearing, the attorney for the party that is at risk (defendant or plaintiff) not infrequently will spend tens (or hundreds) of hours or more, often in consultation with one or more psychologists, either in that community or in a faraway state, in a quest to develop a strategy to attack the bases of the examining psychologist's opinions and thus damage the credibility of the psychologist–consultant whose psychological assessment results and conclusions appear so damaging to his or her client.

Courtroom Questions of Reliability and Validity

The effect has been that increasing numbers of us who practice clinical psychology and clinical neuropsychology (whose knowledge of the bases for the reliability and validity of the most frequently used psychological instruments, including the clinical interview, is usually dated) have had to return to the library in order to better prepare answers to the most searching questions we have been asked since the days we suffered through our doctoral preliminary or final oral examinations. For example, what percentage of us could quote a decade ago—if such data even existed then—the definitive published study that showed that because of its presumed good test–retest stability, an MMPI on a patient would reliably produce exactly or essentially the same general profile in three examinations each about one year apart? And, more to the issue, how many of us could point to research demonstrating such comparability in the three MMPI profiles relevant to the individual case being litigated, and not applicable only to a group of individuals? Or who among us could cite the specific references reporting acceptable studies that show that for a person alleging a severe occupationally induced stress disorder, the MMPI cannot be faked by an individual intent upon doing so, even to the point that the validity scales and overall MMPI profile do not reflect such dissimulation?

For each of today's highly respected psychological tests, other examples of such searching questions abound. However, although my arguments apply to all types of psychological assessment, because of space limitations in this discussion I will deal primarily with issues involved in the assessment of impairment of brain–behavior functions from their higher premorbid level. Thus, most clinical psychologists of my post World War II generation were taught, following Rapaport, Gill, and Schafer (1945), that in clinical situations an individual's premorbid intellectual ability can be determined within an acceptable error range by the person's highest Wechsler subtest score. Furthermore, most of us learned that the lowest subtest scores of a psychotic or head-injured individual may reflect current deficits in the cognitive functions tapped by those subtests with the lowest scores. However, which practitioner among us who has ever been vigorously cross examined on this fairly universally held belief among practicing psychologists regarding premorbid abilities will ever again so nonchalantly assert such an opinion? That is, which practitioner among us is prepared to present evidence acceptable to a jury under the fierce cross examination of an opposing attorney who has been well prepared, for example, by a highly experienced and American Board of Professional Psychology board-certified practitioner, that the difference of 7 points between a Wechsler Adult Intelligence Scale-Revised (WAIS–R; Wechsler, 1981) Digit Span scaled subtest score of 4 and the score of 11 on the Information or Vocabulary subtest, reflects a real difference following a head injury and not a difference that also occurs in the healthy, community-living adult?

We practitioners may vaguely recall from our graduate-student days studying in Wechsler's standardization sample of many hundreds of healthy subjects a table showing the intercorrelation of each Wechsler subtest with every other subtest. And we also may recall that some subtests intercorrelate very high (e.g., WAIS–R Vocabulary correlates .81 with Information), whereas other pairs correlate quite low (Vocabulary correlates only .41 with Object Assembly and only .47 with Digit Symbol). Yet who among us ever expected to be asked the implication of these two low-*r* values by an attorney after we had testified that, *even without other objective corroborative findings,* a seemingly abnormally low score on Digit Symbol and on Object Assembly relative to a relatively high WAIS–R Vocabulary score suggests that the brain–behavior functions involved in executing the Digit Symbol and Object Assembly functions appear impaired relative to the verbal functions associated with Vocabulary, inasmuch as this still high vocabulary score taps functions that are among the most robust indexes of the individual's pre-injury level of neuropsychological functioning? Such opinions were rarely challenged when we presented them during case conferences in a hospital or clinic practice. Yet, as I will elaborate, when millions of dollars are involved in litigation, the meaning of that *r* of only .41 is very clear to the attorney who has used another experienced and well-informed psychologist as a consultant.

We practitioners may also vaguely recall from our graduate-student days the standard error of measurement of a Wechsler subtest score and may even more vaguely recall that somehow it also was a useful index for determining the probability that an obtained difference in two subtest scores for an individual was a statistically valid difference. But on the witness stand, under intense cross examination, who among us can draw out the full implication of such a standard error in reference to the clinical opinion or conclusion regarding this individual patient that we have just offered to the jury?

The same types of questions apply to our other psychological tests. For example, who among psychologist–practitioners on the witness stand can easily recall the intercorrelations (and their implications for the case being litigated) of each subtest in the Halstead-Reitan Battery with every other subtest? Or, who can recall whether Ward Halstead, Ralph Reitan, or any other investigator ever published the standard error of measurement of each of their tests? There is little that is more humbling to a practitioner who uses the highest one or two Wechsler subtest scores as the only index of a patient's "premorbid" level of intellectual functioning and who therefore interprets concurrently obtained lower subtest scores as indexes of clear "impairment" and who is then shown by the opposing attorney elementary and high school transcripts that contain several global IQ scores, each of which were at the same low IQ levels as are suggested by the currently obtained lowest Wechsler subtest scaled scores.

What have I and many other practitioners (but unfortunately experience suggests not all) learned from these questions and similar grillings during the past 15 years? Quite a bit. Society has accorded professional psychologists the privileged status of expert witness. As such we are involved in human dramas and in decisions that are extremely costly not only to the humans involved but also to insurance companies and other segments of society, which pay the millions of dollars juries award, often because of the expert testimony psychologists have contributed. I have reached the opinion that the intensive examination of the validity of our clinical opinions in the courtroom will motivate psychologists to improve even further the validity of our work in psychological assessment, rather than interfere with the quality of such work.

Given these remarks, those who are familiar with my writings will not be surprised when I acknowledge that much of my research in clinical neuropsychology during the last 15 years (which has been focused on the reliability and validity of the instruments such as the adult Wechsler scales and the tests that make up the Halstead-Reitan Neuropsychological Battery [HRB] that you and I use as aids to our clinical judgment) was largely motivated by the grillings I have endured on the witness stand. Although a considerable body of such research has been published by others, I will now present a few highlights from recent research of mine that bears on the validity of the opinions some of you and I are being asked to give in the courtroom.

Premorbid Cognitive Functioning: Relevant Psychometric Indexes

In current personal injury litigation, one of the questions most frequently asked of clinical neuropsychologists by attorneys for both the plaintiff and the defense is whether the person who experienced the accident sustained a brain injury, and if so, which particular brain–behavior functions were impaired and how much. In my experience, in annually increasing numbers of such cases *that actually reach the courtrooms,* the findings from comprehensive medical and neurological examinations (which include laboratory studies, X-rays, and sophisticated neuro-imaging techniques) reveal no objective evidence of trauma to the brain. Typically, in such cases the only evidence presented by the plaintiff's or defendant's attorney to the consulting clinical neuropsychologist that the patient is psychologically impaired consists of the latter's exceedingly difficult-to-confirm subjective report of headaches, dizziness, confusion, trouble with concentration, and memory, as well as symptoms of depression and anxiety. In these increasing numbers of instances of a total absence of objective findings (from hospital, medical, and neurological records) that would tend to validate the presence of a brain injury, my experience is that the opinions proffered by too many clinical-psychologist–examiners unfortunately are based *solely* on the data gathered during the latters' examination of the patient. That is, they are opinions based only on the psychological test scores and arrived at without studying, let alone integrating, those test scores with objective information contained in the personal–social history (e.g., school transcripts, military records, pertinent information in the patient's job-related personnel records). Those records often provide highly useful data with which to establish a pre-injury baseline against which to compare the reported subjective symptoms and the findings from the current neuropsychological examination (Matarazzo, 1972; Matarazzo & Herman, 1984a, 1984b, 1985). For example, even when all medical and hospital findings are negative, the finding of a current WAIS-R Full-Scale IQ (FSIQ) of 85, with comparable low scores on the Wechsler Memory Scale–Revised (WMS–R; Wechsler, 1987) and Halstead-Reitan Battery, in a patient whose school transcripts record actual IQ and SAT scores in the above-average range, buttressed by a compatible life history (i.e., accountant), provide useful and persuasive evidence by which to evaluate the cognitive losses associated with the patient's subjectively experienced symptoms of headache, confusion, memory impairment, and so on.

In addition, a survey of the literature (Leckliter & Matarazzo, 1989) provides ample evidence that in healthy individuals such variables as age, education, IQ, gender, and alcohol abuse markedly influence scores on the neuropsychological tests included in the Halstead-Reitan Battery. Figure 1, constructed from Table 3 of Leckliter and Matarazzo, presents in summary form the findings for one such variable. Experience in the courtroom suggests that some psychologists are unfamiliar with the effects on healthy individuals of such influences, and erroneously conclude that "impairment" due to an alleged recent brain trauma is present when, in fact, the abnormal-appearing scores are due only to normal advancing age; or a limited number of years of prior education; or documented, lifelong, substantially below-average indexes of measurable intelligence; or many years of alcohol abuse; and so on.

Published information on the psychometric properties of the tests used in currently administered clinical neuropsychological examinations, along with information from the school and work records, is too infrequently used by some practitioners. Such information on psychometric properties is critically important for the clinician who is interpreting test scores in order to reach an assessment conclusion that is anchored in the person's life history. For example, published information on the following test properties is critical, albeit not used enough, when discerning whether the psychological test results are consistent with a conclusion of impairment relative to the patient's inferred higher level of cognitive functioning: the test's mean, standard deviation, standard error of measurement, and test–retest reliability, in addition to relevant tables of test (or subtest) intercorrelations. An appreciation of the implication in clinical practice of such psychometric properties would do much to decrease the numbers of patients being diagnosed as showing cognitive impairment based almost exclusively on their subjective symptoms (e.g., some of those alleging exposure to neurotoxins in a work environment judged to be safe by state inspectors) plus the findings of some high and low scores when administered a battery of neuropsychological tests. Although I will provide primarily such psychometric information for the subtests that make up the WAIS–R, the data for their computation or the actual values of the comparable psychometric properties of the MMPI, HRB, Wechsler Memory (WMS–R), and most other tests currently in use also have been published and are critically important in clinical practice.

Standard Error of Measurement of a Test Score

For the individuals in the reference group upon whom the WAIS–R was standardized, the mean of each of the 11 individual subtests was set at 10 with a standard deviation of 3. Likewise, for each age group, the mean Verbal IQ (VIQ), Performance IQ (PIQ), and FSIQ was set at 100, and the standard deviation of each of the three was set at 15 (Wechsler, 1981, pp. 24–25). Fortunately, most clinicians are aware of these properties of the WAIS–R. However, relative to the attacks in the courtroom on the validity of our test-based interpretations, the standard errors of each of the 11 subtests as well as each of the three IQ scores are much more important. Specifically, the magnitudes of those standard errors of an obtained scaled score on the 11 individual subtests range from 0.61 of a scaled score point for the Vocabulary subtest to 1.54 points for the Object Assembly (OA) subtest. Likewise, the standard errors of an obtained VIQ, PIQ, and FSIQ are 2.74, 4.14, and 2.53 scaled score points, respectively (Wechsler, 1981, pp. 31–34).

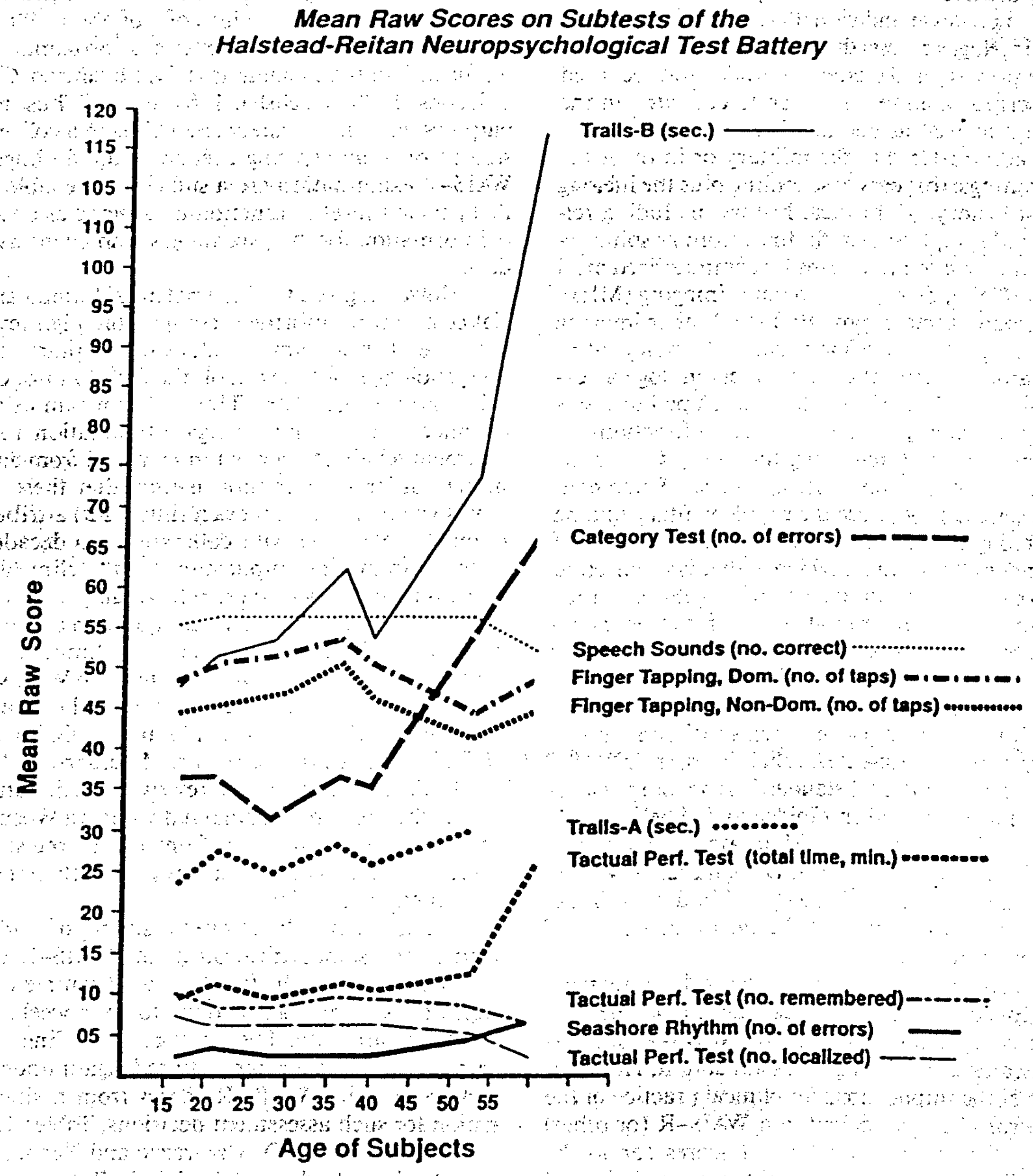
It is important for clinicians to recall that the magnitudes of these standard errors of measurement indicate the actual band of error around each obtained IQ or each subtest scaled score; and thus they highlight the risk of interpreting such a score (from only a single WAIS–R examination) as constituting an exact quantitative index of an underlying brain–behavior attribute of that individual (e.g., the patient's presumed premorbid level of cognitive functioning or a current impairment from that earlier level).

When the goal is to infer whether functioning is impaired, a good method of taking into account the standard error of measurement is for the practitioner to consider a band of scores extending two standard errors above and two below the obtained score. Thus, an obtained OA score of 9 communicates that given the other-than-perfect reliability of such a test score that was sampled only once, the practitioner may be confident at the .05 level that the patient's true OA score is probably not 9 but, instead, falls between 5.92 and 12.08 (i.e., plus and minus 2 times the standard error of 1.54 points of the obtained score of

9); and at the .01 level falls between 5.00 and 13.00 (i.e., plus and minus 2.6 times that standard error). Comparably, an obtained WAIS–R FSIQ of 100 indicates that the practitioner may be confident that at the .05 level the patient's true FSIQ falls between 94.94 and 105.06; and at the .01 level of statistical significance (and thus even higher confidence) the FSIQ falls between 93.42 and 106.58. Use by the clinician in these several calculations of the obtained subtest scaled score and FSIQ score, rather than the patient's true OA and FSIQ scores, although not completely accurate psychometrically (Dudek, 1979), nevertheless provides what Gulliksen (1950) called the "reasonable limits" of each of these two true scores.

My experience in the courtroom, where more and more psychologists' conclusions are being vigorously challenged by attorneys, has led me to conclude that too many psychologists testifying in the courts today, whether for the plaintiff or defense, are unaware of the standard errors of measurement of the scores (and accompanying confidence intervals) produced by our batteries of tests. Unfortunately, *in those cases that actually reach the courtroom*, today's modal pattern (in conclusions related

Figure 1
The Relationship of Halstead-Reitan Test Scores and Age in Healthy Control Subjects



to higher premorbid functioning) is to diagnose impairment in brain functioning exclusively from a patient's subjectively reported symptoms plus one or more low scores in a battery of tests that included the WAIS–R and the HRB subtests. Because of the magnitudes of their standard errors, in isolation and without other objective corroborating evidence, a finding of a sizeable degree of *scatter* (subtest-to-subtest differences) in a set of WAIS–R and HRB scores cannot be used ipso facto to either (a) estimate (using the highest scores) the examinee's supposed premorbid level of cognitive function or (b) identify areas (using the lowest scores) of current cognitive impairment.

However, clinical judgments and conclusions such as these two are possible when such high and low subtest scores are evaluated in a more comprehensive clinical context. This includes an individual's (a) premorbid scores obtained on intelligence tests that were administered years earlier in the primary and secondary grades and recorded on the transcripts of almost every child educated in the United States, as well as comparable tests of cognitive functioning administered in the military or in other occupational settings; (b) years of schooling plus the lifelong occupational history; (c) medical history, including relevant signs and symptoms; (d) findings from hospital records, including one or more scans by computerized axial tomography (CAT), magnetic resonance imaging (MRI), positron emission tomography (PET) and other imaging procedures; and (e) other relevant supplementary information obtained during the current psychological examination from the use of other tests developed to assess related neuropsychological and personality functions.

Caution is in order regarding the use of the results obtained with other neuropsychological tests. Some neuropsychologists exclusively use the supplementary finding of scatter among the subtests of the HRB as examples of this independent "objective" evidence that is required to help one conclude that scatter on the WAIS–R is mirroring an impairment in brain–behavior function. However, that subtest-to-subtest scatter is as common in normal individuals across subtests of the HRB, WMS–R, and related batteries as it is across subtests of the WAIS–R, may be inferred in part from such tests' comparable (a) other-than-perfect test–retest reliabilities, (b) tables of subtest intercorrelations, (c) standard errors of measurement (which are reported in Goldstein & Shelly, 1971, 1972; Halstead, 1947; J. D. Matarazzo, Matarazzo, Wiens, Gallo, & Klonoff, 1976; J. D. Matarazzo, Wiens, Matarazzo, & Goldstein, 1978; Royce, Yeudall, & Bock, 1976; Wechsler, 1987), and (d) the factor structures of the HRB and other tests.

When used as only one of a number of documentable indexes of loss of earlier intellectual capacity of the type that accompanies brain impairment, WAIS–R subtest-to-subtest scatter can be a highly useful datum. However, knowledge of the implications for clinical practice of the standard error of measurement of a WAIS–R (or other) test score makes clear why such test scores cannot be used in isolation for diagnosing either an individual's higher premorbid level of cognitive functioning (from the highest subtest score) or to identify (using the lowest subtest scores) areas of current cognitive impairment.

I will now describe a few other well-known psychometric properties of psychological test scores that practitioners appear to be neglecting in the opinions they are offering in courtroom testimony involving personal injury litigation relating to brain impairment. As I indicated earlier, my interest in carrying out the research was stimulated by my courtroom experience during the past decade.

Test–Retest Reliability of a Test Score

The magnitudes of the test–retest reliabilities of the scores yielded by intelligence tests such as the WAIS–R are well-known to psychology students as well as to practitioners, namely, retest stability values of *r* of about .90 for each of the three IQ scores and between approximately .70 to .90 for each of the 11 subtests (J. D. Matarazzo, Carmody, & Jacobs, 1980; Wechsler, 1981, p. 32). Thus, for many purposes (e.g., in the career counseling of a college-bound student or of an aspiring executive) the findings from a WAIS–R examination are a sufficiently reliable index of that person's level of functioning to serve as a valid item of information for the psychologist and client using such data.

However, given the high potential human and financial costs, an uninformed, comparably high level of acceptance of these same *r* values in the practice of neuropsychology would overlook the fact that none of these values is at or near 1.00. Thus, the clinician using scores obtained during only a single examination to reach a judgment relating to current impairment from an inferred higher earlier level cannot assume that these obtained scores validly mirror an exact (invariant) attribute of the examinee. Working with colleagues two decades ago to better examine the implication for the clinical practice of neuropsychology of these other-than-perfect test–retest values of *r*, I began to pursue the question of the potential error in inferring "impairment" or "improvement" or arriving at related clinical decisions from Wechsler scores obtained from a single examination (J. D. Matarazzo et al., 1980; R. G. Matarazzo, Matarazzo, Gallo, & Wiens, 1979; R. G. Matarazzo, Wiens, Matarazzo, & Manaugh, 1973). That research, plus a review of the literature, made it clear that for some normal individuals a Wechsler score could change significantly from test to retest and thus could not be considered an invariant attribute for such assessment conclusions.

Concurrently, data from the sample of 1,880 subjects used in the standardization of the WAIS–R (Wechsler, 1981) included for the first time a subsample of 119 subjects who were re-examined five to seven weeks after their initial examination. These test-retest findings, which we shortly thereafter further analyzed, again underscore the dangers of using WAIS–R scores from a single examination for such assessment decisions. Tables 1, 2, and 3, reproduced from J. D. Matarazzo and Herman (1984b), present those findings. The WAIS–R findings in these

Table 1

Frequency of Different Magnitudes of Gain or Loss in Verbal IQ, Performance IQ, and Full Scale IQ From Initial Test to Retest for 119 Adults in the WAIS–R Standardization Sample

Gain or loss	Verbal IQ		Performance IQ		Full Scale IQ	
	n	%	*n*	%	*n*	%
−12	1	0.8	1	0.8	1	0.8
−11	0	0.0	0	0.0	0	0.0
−10	1	0.8	0	0.0	0	0.0
−9	0	0.0	0	0.0	0	0.0
−8	0	0.0	1	0.8	0	0.0
−7	0	0.0	0	0.0	0	0.0
−6	0	0.0	0	0.0	0	0.0
−5	3	2.5	2	1.7	1	0.8
−4	3	2.5	1	0.8	2	1.7
−3	4	3.4	1	0.8	2	1.7
−2	4	3.4	1	0.8	1	0.8
−1	8	6.7	3	2.5	1	0.8
0	7	5.9	3	2.5	5	4.2
1	7	5.9	8	6.7	6	5.0
2	15	12.6	5	4.2	7	5.9
3	10	8.4	7	5.9	7	5.9
4	10	8.4	9	7.6	13	10.9
5	9	7.6	4	3.4	6	5.0
6	4	3.4	7	5.9	9	7.6
7	13	10.9	4	3.4	8	6.7
8	8	6.7	6	5.0	7	5.9
9	4	3.4	4	3.4	13	10.9
10	1	0.8	6	5.0	8	6.7
11	2	1.7	5	4.2	9	7.6
12	2	1.7	4	3.4	3	2.5
13	2	1.7	4	3.4	3	2.5
14	0	0.0	12	10.1	2	1.7
15	1	0.8	3	2.5	1	0.8
16	0	0.0	4	3.4	1	0.8
17	0	0.0	1	0.8	1	0.8
18	0	0.0	2	1.7	0	0.0
19	0	0.0	1	0.8	1	0.8
20	0	0.0	3	2.5	1	0.8
21	0	0.0	2	1.7	0	0.0
22	0	0.0	0	0.0	0	0.0
23	0	0.0	1	0.8	0	0.0
24	0	0.0	1	0.8	0	0.0
25	0	0.0	0	0.0	0	0.0
26	0	0.0	1	0.8	0	0.0
27	0	0.0	1	0.8	0	0.0
28	0	0.0	1	0.8	0	0.0
Total	119	100	119	100	119	100
Mean	3.3		8.4		6.2	

Note. WAIS–R = Wechsler Adult Intelligence Scale–Revised. From "Base-Rate Data for the WAIS–R: Test–Retest Stability and VIQ–PIQ Differences" by J. D. Matarazzo and D. O. Herman, 1984, *Journal of Clinical Neuropsychology, 6,* p. 354. Copyright 1984 by the Psychological Corporation. Reprinted by permission.

three tables, consistent with what we had published in 1973 for the WAIS, reveal in another form the information communicated by the standard error of measurement of the test scores I described earlier. Namely, although each of the 3 IQ or 11 subtest scaled scores from the WAIS–R is a fairly reliable index of an individual's level of functioning, such a score, obtained *solely* from one examination, will on re-examination be the same or approximately the same in many instances, but will be quite a bit different (higher or lower) in other cases. Clearly, then, when interpreting the WAIS–R score of an individual who is examined only once, a practitioner cannot assume that such a score exactly represents the true magnitude of the associated attribute being measured. Specifically, the findings in Tables 1, 2, and 3 indicate that because of measurement error, a second examination (or an infinite number of examinations) of that attribute (IQ or scaled subtest score) of that individual, with the same or good alternate forms of that test, will inevitably produce some higher as well as some lower scores for that very same attribute. Consequently, such an IQ or subtest score cannot be used in isolation or be accompanied only by the subjective report of the examinee in reaching a diagnostic conclusion regarding "enhanced" or "impaired" cognitive functioning.

Scatter in WAIS–R Subtest Scores

A similar, clinically important conclusion follows from a subsequent set of findings that we reported from another analysis of the WAIS–R standardization data (J. D. Matarazzo, Daniel, Prifitera, & Herman, 1988; J. D. Matarazzo & Prifitera, 1989). The purpose of that study was to examine the degree to which subjects from that representative normal sample show variability in the magnitudes of their own scores across the 11 subtests of the WAIS–R (i.e., for any given subject the difference, or scatter, in points between his or her highest and lowest subtest score). For the 1,880 subjects, the results shown here in Tables 4 and 5 lead to a sobering conclusion: Even when it is substantial, such scatter is by itself *not* an indicator of brain dysfunction, inasmuch as it is a characteristic of the cognitive functioning of normal[1] subjects. Specifically, as shown in Table 4, in normal subjects the average difference between the highest and lowest WAIS–R subtest scaled score was 4.67 (with a range for any given individual of 2–13 points) across the 6 Verbal subtests, 4.71 (with a range of 1–16 points) across the 5 Performance subtests, and 6.66 (with a range of 2–16 points) across the same 11 subtests in the Full Scale. Table 5 elaborates on that finding.

Although the test–retest reliability (Tables 1, 2, and 3), as well as the errors of measurement associated with each WAIS–R score discussed earlier, plus the data shown in Tables 4 and 5, collectively highlight the risk associated with a clinician's determining premorbid IQ solely from an individual's highest subtest scores, the data in the

[1] In the selection of the 1,880 subjects for the Wechsler Adult Intelligence Scale–Revised standardization sample, examiners were asked to not include individuals with known brain damage, severe behavioral or emotional problems, physical defects that would restrict their ability to respond to test items, or who could not speak and understand English (Wechsler, 1981, p. 18).

Table 2
Frequency of Different Magnitudes of Gain or Loss in Scaled Scores on the Verbal Subtests From Test to Retest For 119 Adults in the WAIS–R Standardization Sample

Gain or	INF		DSP		VOC		ARITH		COMP		SIM	
loss	n	%	n	%	n	%	n	%	n	%	n	%
−9												
−8												
−7												
−6	0		0		0		0		1	0.8	0	
−5	0		1	0.8	0		0		1	0.8	0	
−4	1	0.8	0	0.0	0		0		2	1.7	1	0.8
−3	1	0.8	3	2.5	2	1.7	2	1.7	3	2.5	0	0.0
−2	2	1.7	9	7.6	6	5.0	11	9.2	9	7.6	3	2.5
−1	10	8.4	14	11.8	17	14.3	12	10.1	22	18.5	23	19.3
0	50	42.0	37	31.1	50	42.0	36	30.3	34	28.6	33	27.7
1	35	29.4	26	21.8	29	24.4	24	20.2	21	17.6	20	16.8
2	12	10.1	18	15.1	10	8.4	18	15.1	11	9.2	18	15.1
3	5	4.2	9	7.6	5	4.2	11	9.2	10	8.4	9	7.6
4	1	0.8	2	1.7	0	0.0	3	2.5	4	3.4	9	7.6
5	2	1.7	0	0.0	0	0.0	1	0.8	0	0.0	2	1.7
6	0	0.0	0	0.0	0	0.0	1	0.8	1	0.8	1	0.8
7												
8												
9												
Total	119	100	119	100	119	100	119	100	119	100	119	100
Mean	0.6		0.4		0.2		0.6		0.2		0.9	

Note. WAIS–R = Wechsler Adult Intelligence Scale–Revised. INF = information; DSP = Digit Span; VOC = Vocabulary; ARITH = Arithmetic; COMP = Comprehension; SIM = Similarities. From "Base-Rate Data for the WAIS–R: Test–Retest Stability and VIQ–PIQ Differences" by J. D. Matarazzo and D. O. Herman, 1984, *Journal of Clinical Neuropsychology, 6,* p. 356. Copyright 1984 by the Psychological Corporation. Reprinted by permission.

WAIS–R standardization sample provides information in still a different form that neuropsychologists should also find persuasive. Specifically, to display each range of scatter—using as our only criterion that they each fall in a designated range—we randomly selected 20 protocols from among the 1,880 normal individuals in the standardization sample. Table 6 presents the scores on each of the 11 WAIS–R subtests for each of these 20 representative individuals whose person-specific amounts of scatter between their highest and lowest subtest scores were 3, 4, 6, 7, 8, 10, 12, 13, and 15 points. (Table 5 shows how common each of these 9 selected magnitudes of scatter were among the 1,880 individuals.) As shown in Table 6, these 20 individuals included 9 women and 11 men whose ages ranged from 16 to 64 and whose FSIQs ranged from 74 to 131. The data in Table 6 leave little question that substantial scatter is characteristic of normal subjects, and thus, in isolation, scatter tells the clinician nothing about impairment of the type also shown by patients in whom other objective evidence confirms the presence of a brain injury. A fuller discussion of the implications for the practitioner of the findings on scatter shown in Tables 4, 5, and 6 is included in the original publications (Matarazzo et al., 1988; Matarazzo & Prifitera, 1989).

It should not be inferred from the preceding discussion that because a substantial magnitude of scatter is common in normal records, it follows that such scatter across a battery of tests is never clinically informative. In fact, even when all of the medical and neurological findings are negative, interpreted within the context of information in the personal and social history, there may be times when a difference of a few scaled points between two WAIS–R subtests is clinically meaningful and requires further analysis. An example would be scores of 12 on Arithmetic and 9 on Vocabulary earned by a widely published professor of English. Conversely, a review of personal, educational, medical, and hospital records and neuropsychological test findings may not clinically support the inference of impairment for an individual with scatter of 8 points or more on the Verbal subtests (e.g., a Vocabulary score of 19 and an Arithmetic score of 11 for the English professor). Obviously, in such cases clinical experience and informed judgment continue to play an irreplaceable role.

When integrated with both the personal and social history and relevant medical and clinical findings, the amount of subtest scatter may be highly significant. Detailed findings in three clinical cases in which scatter found across different subtests proved useful is described in Matarazzo (1972, pp. 414–427). A detailed summary

Table 3

Frequency of Different Magnitudes of Gain or Loss in Scaled Scores on the Performance Subtests From Test to Retest For 119 Adults in the WAIS–R Standardization Sample

Gain or loss	PC		PA		BD		OA		DSY	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
−9										
−8										
−7										
−6										
−5	0		1	0.8	0		1	0.8	0	
−4	0		1	0.8	0		2	1.7	0	
−3	3	2.5	4	3.4	2	1.7	0	0.0	2	1.7
−2	2	1.7	3	2.5	6	5.0	3	2.5	4	3.4
−1	5	4.2	15	12.6	13	10.9	7	5.9	9	7.6
0	39	32.8	21	17.6	32	26.9	21	17.6	30	25.2
1	25	21.0	25	21.0	41	34.5	21	17.6	41	34.5
2	26	21.8	14	11.8	13	10.9	17	14.3	14	11.8
3	14	11.8	11	9.2	8	6.7	21	17.6	14	11.8
4	2	1.7	12	10.1	2	1.7	11	9.2	2	1.7
5	2	1.7	6	5.0	2	1.7	8	6.7	2	1.7
6	1	0.8	5	4.2	0	0.0	5	4.2	0	0.0
7	0	0.0	0	0.0	0	0.0	1	0.8	1	0.8
8	0	0.0	1	0.8	0	0.0	1	0.8	0	0.0
9										
Total	119	100	119	100	119	100	119	100	119	100
Mean	1.1		1.3		0.7		1.9		0.9	

Note. WAIS–R = Wechsler Adult Intelligence Scale–Revised. PC = Picture Completion; PA = Picture Arrangement; BD = Block Design; OA = Object Assembly; DSY = Digit Symbol. From "Base-Rate Data for the WAIS–R: Test–Retest Stability and VIQ–PIQ Differences" by J. D. Matarazzo and D. O. Herman, 1984, *Journal of Clinical Neuropsychology, 6,* p. 357. Copyright 1984 by the Psychological Corporation. Reprinted by permission.

Table 4

Average Difference (Scatter) Between an Individual's Highest and Lowest Subtest Scaled Score: Data From the Three Scales of the WAIS–R Standardization Sample

	IQ range					
Scale	79 and under	80–89	90–109	110–119	120+	Total sample
Verbal						
Mean scatter	3.48	4.05	4.75	5.28	5.35	4.67
Range	2–8	2–10	2–12	2–13	2–10	2–13
Performance						
Mean scatter	3.36	4.32	4.81	5.05	5.53	4.71
Range	2–11	1–15	2–16	2–14	2–13	1–16
Full						
Mean scatter	5.02	5.93	6.85	7.15	7.65	6.66
Range	3–11	2–12	3–16	4–15	4–13	2–16
n	165	302	924	312	177	1,880

Note. WAIS–R = Wechsler Adult Intelligence Scale–Revised. From "Inter-subtest Scatter in the WAIS–R Standardization Sample" by J. D. Matarazzo, M. H. Daniel, A. Prifitera, and D. O. Herman, 1988, *Journal of Clinical Psychology, 44,* pp. 945, 946, 947. Copyright 1989 by the Psychological Corporation. All rights reserved. Reprinted by permission. Also, from "Subtest Scatter and Premorbid Intelligence: Lessons From the WAIS–R Standardization Sample" by J. D. Matarazzo and A. Prifitera, 1989, *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1,* p. 188. Copyright 1989 by the American Psychological Association. Reprinted by permission.

of a fourth individual who showed clinically significant scatter was published more recently (Matarazzo, 1985, p. 250). It is important to emphasize that the amount of scatter shown in the scores of those four patients, including a substantial difference between each individual's VIQ and PIQ in three of the four, was clinically important

Table 5
Full Scale: Percentage of Cases at or Above Each Magnitude of Scatter Across the Full Scale for the 1,880 Subjects in the WAIS–R Standardization Sample

Scatter: Difference in points between highest and lowest of 11 subtest scaled scores	Percentage of cases showing this or more points of scatter	No. of individuals showing this magnitude of scatter
17	0.0	0
16	0.1	2
15	0.3	4
14	0.4	2
13	1.0	11
12	2.1	20
11	4.1	38
10	8.6	84
9	18.1	180
8	31.9	258
7	48.7	316
6	69.1	384
5	86.1	320
4	96.5	195
3	99.6	58
2	99.9	7
1	100.0	1
0	100.0	0

Note. Mean scatter = 6.66 (*SD* = 2.08). Median scatter = 6. WAIS–R = Wechsler Adult Intelligence Scale–Revised. From "Inter-subtest Scatter in the WAIS–R Standardization Sample" by J. D. Matarazzo, M. H. Daniel, A. Prifitera, and D. O. Herman, 1988, *Journal of Clinical Psychology, 44*, p. 945. Copyright 1989 by the Psychological Corporation. All rights reserved. Reprinted by permission. Also, from "Subtest Scatter and Premorbid Intelligence: Lessons From the WAIS–R Standardization Sample" by J. D. Matarazzo and A. Prifitera, 1989, *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1*, p. 188. Copyright 1989 by the American Psychological Association. Reprinted by permission.

because of its relation to other objective, corroborating information. This point needs special emphasis inasmuch as in another study with the WAIS–R standardization sample, Matarazzo and Herman, 1984b, 1985, found that the differences between VIQ and PIQ for each of these 1,880 subjects, although normally distributed around a mean VIQ–PIQ difference of zero points, actually showed a remarkably large standard deviation of 11.12 and a range of −43 to +49 points across these individual normal subjects.

Once again, even very marked VIQ–PIQ scatter is not necessarily pathognomonic when it constitutes the *only* evidence of impairment. I now turn to another psychometric property of tests with which the practitioner must be familiar.

Intercorrelations of Tests and Subtests of Cognitive Functioning

Tables of the (inter)correlations of each of the 11 subtests with each other subtest have been published since 1939 for each of Wechsler's intelligence scales. However, only during the last decade did I realize that the implications of the data in those tables were unclear to me and to my generation of clinicians whose earlier practices involved our own patients seen almost exclusively in clinical settings, not patients referred by attorneys. As the numbers of patients involved in personal injury litigation (and for whom we had no continuing responsibility) increased, the courtroom requirement that we more impersonally involved practitioners better defend our opinions forced me to reacquaint myself with this additional psychometric property of the tests we use. Specifically, Table 7, taken from Wechsler (1981, p. 46), shows the intercorrelation of scores on each of the 11 WAIS–R subtests for all 1,880 subjects in the standardization sample. Standing alone, these subtest-to-subtest correlations, as well as the correlations presented in rows 4 to 6 from the bottom of Table 7, again undermine the two-pronged thesis that an individual's highest subtest scores validly reflect premorbid level of intelligence and that the lowest scores mirror impaired functions. Specifically, as shown in the fourth row of numbers from the bottom, the fact that the correlations between FSIQ and the scores on each of the 11 subtests are far from unity (ranging only from .57 to .81) strongly indicates that, used in isolation, no single subtest score (or combination thereof) is an acceptable measure of a normal person's (let alone a patient's) presumed actual level of (premorbid) FSIQ. In addition, the data in the body of Table 7 reveal that whereas scores on some pairs of subtests show an acceptably high correlation (i.e., the score on the Vocabulary subtest correlates .81 with the score on the Information subtest), the correlation across other pairs of subtests is unacceptably low, even in normal subjects (i.e., the score on the Digit Symbol subtest correlates only between .38 and .47 with the score on each of the other 10 subtests), to permit using high and low subtest scores as a diagnostic means for ascertaining impairment.

Although their length precludes my reproducing them here, whether the WAIS–R subtests are intercorrelated alone or are combined and intercorrelated with the subtests of other batteries such as the subtests of the HRB, the WMS–R, and other batteries of neuropsychological tests (Goldstein & Shelly, 1971, 1972; Royce et al., 1976; Wechsler, 1987), the resulting tables of intercorrelations contain a range in values of correlations that are like those shown in Table 7 for the WAIS–R. Once again, the fact that only a few of these correlations in such tables approach unity means that a mix of both high and low subtest scores is the norm, even in the unimpaired, healthy individual.

One conclusion is clear from Table 7, as well as from these just-cited, expanded tables: Use of high and low subtest scores in the WAIS–R, HRB, and WMS–R for determining either premorbid ability or impairment, in isolation and without corroboration using the types of independent information described earlier, is unjustifiable. My experience to date suggests that it is only a matter of time before more plaintiff and defense attorneys will incorporate into their questions the meaning of the findings in such tables of intercorrelation as the one in Table 7.

Table 8
WAIS–R Standardization Sample: The 11 Subtest Scaled Scores of 20 Representative Individuals Showing Differences (Scatter) From 3 to 15 Points

				Subtest scaled scores[a]										
Range of scatter	FSIQ	Sex	Age	INF	DS	VOC	ARITH	COMP	SIM	PC	PA	BD	OA	DSY
3	76	F	70–74	5	6	5	3	4	3	3	4	5	3	3
3	86	F	55–64	5	8	8	8	8	5	5	5	6	8	5
3	104	M	25–34	12	11	11	12	10	12	10	9	12	12	10
4	116	M	35–44	14	11	12	12	12	10	11	11	11	12	12
4	122	F	45–54	12	12	12	13	11	14	12	12	12	12	10
6	77	F	25–34	8	9	8	6	6	9	3	7	4	5	8
6	85	M	25–34	10	10	11	10	8	7	8	6	5	5	8
6	100	M	65–69	10	11	10	8	9	10	8	6	7	6	5
7	111	F	70–74	12	9	12	6	10	9	10	6	8	9	5
7	131	M	16–17	12	13	12	15	13	14	12	9	11	9	16
8	81	M	65–69	10	8	8	6	6	2	6	2	2	3	2
9	74	M	25–34	5	4	4	4	3	6	9	6	5	12	6
9	95	F	45–54	14	6	11	7	12	9	10	7	6	5	6
10	87	F	25–34	9	8	7	9	6	8	10	9	6	5	15
10	115	M	25–34	16	6	15	13	13	11	12	11	12	13	12
10	123	F	55–64	15	8	13	9	18	10	8	15	8	11	13
12	101	F	18–19	6	8	9	7	13	9	10	17	9	9	5
13	116	M	45–54	16	11	9	11	4	14	12	6	17	13	12
13	131	M	70–74	16	14	19	12	13	15	6	10	10	6	6
15	96	M	35–44	11	2	11	9	7	10	14	11	1	8	16

Note. WAIS–R = Wechsler Adult Intelligence Scale–Revised; FSIQ = Full-Scale IQ; F = female; M = male; INF = Information; DS = Digit Span; VOC = Vocabulary; ARITH = Arithmetic; COM = Comprehension; SIM = Similarities; PC = Picture Completion; PA = Picture Arrangement; BD = Block Design; OA = Object Assembly; DSY = Digit Symbol. Data in this table are from the WAIS–R standardization sample. Copyright 1989 by the Psychological Corporation. All rights reserved. Reprinted by permission.

[a] Using the scaled-score conversions for the reference group (ages 20–34).

Cognitive Functions: Unitary or Highly Differentiated

The tables of intercorrelations just described fail to support another widely held belief that is becoming increasingly evident in the reports of many neuropsychology practitioners, namely, the belief that the different individual subtests of batteries such as the WAIS–R, WMS–R, and HRB validly assess brain-area-related, clear-cut, functional differences in cognition-specific intellectual, memory, constructional, motor, orientation, attentional, executive, and other so-called discrete and highly differentiated neuropsychological functions. Alas, the results of factor analyses carried out on the data in the just-cited tables of intercorrelations of tests that make up today's neuropsychological batteries (Goldstein, 1984; Heilbronner, Buck, & Adams, 1989; Kupke & Lewis, 1989; Leckliter & Matarazzo, 1986; Matarazzo, 1972; Royce et al., 1976; Swiercinsky, 1979) reveal that, just as debated by Charles Spearman and E. L. Thorndike almost a century ago (see Matarazzo, 1972; pp. 47–50, and Matarazzo & Prifitera, 1989, pp. 188–190), none of the tests in such batteries has been found to be primarily a measure of one or another of these just-enumerated discrete brain–behavior functions. Rather, these factor analytic studies reveal that each such test is in the main primarily a measure of a common, general cognitive attribute, Spearman's g (Silverstein, 1985), with the rest of the considerably smaller variance probably attributable at the most to two or three considerably weaker group factors. That is, considerably weaker attributes that mirror individual differences in what appear to be second-order group or specific factors such as verbal comprehension, perceptual organization, and sense-specific memory capacity. Consequently, clinical neuropsychologists who from their evaluation describe a patient's strengths and impairments in as many as a dozen and more such allegedly different first or second order, cognition-specific functions as those just enumerated simply are unaware of the clinically highly relevant implications contained in tables of intercorrelations such as our Table 7, or are actually clear from the numerous published factor analyses of such batteries of neuropsychological tests as I just cited.

Psychological Diagnosis and Psychological Assessment

To this point I have discussed the practice of psychological assessment, an activity by which the clinician integrates test findings with information from the personal, educational, and occupational histories as well as from the findings of other clinicians. It should be clear that the portrait of an individual presented in such a typical 10–20-page report, whether accurate or not, is very different from the portrait communicated by a one- or two-word

Table 7
WAIS–R Standardization Sample (N = 1,880): Average Intercorrelation of the Tests for Nine Ages

Test	INF	DS	VOC	ARITH	COMP	SIM	PC	PA	BD	OA	DSY	Verbal score	Performance score
Information (INF)	—												
Digit Span (DS)	.46	—											
Vocabulary (VOC)	.81	.52	—										
Arithmetic (ARITH)	.61	.56	.63	—									
Comprehension (COMP)	.68	.45	.74	.57	—								
Similarities (SIM)	.66	.45	.72	.56	.68								
Picture Completion (PC)	.52	.37	.55	.48	.52	.54	—						
Picture Arrangement (PA)	.50	.37	.51	.46	.48	.50	.51	—					
Block Design (BD)	.50	.43	.52	.56	.48	.51	.54	.47	—				
Object Assembly (OA)	.39	.33	.41	.42	.40	.43	.52	.40	.63	—			
Digit Symbol (DSy)	.44	.42	.47	.45	.44	.46	.42	.39	.47	.38	—		
Verbal score[a]	.79	.57	.85	.70	.76	.74	.61	.57	.61	.49	.54	—	
Performance score[b]	.62	.50	.65	.62	.61	.64	.65	.56	.70	.62	.52	.74	—
Full-Scale score[c]	.76	.58	.81	.72	.74	.75	.67	.61	.68	.57	.57	—	—
Average correlation of tests with Verbal, Performance, and Full-Scale scores before correction for contamination													
Verbal score[a]	.86	.69	.90	.80	.84	.83	—	—	—	—	—	—	—
Performance score[b]	—	—	—	—	—	—	.79	.73	.82	.77	.70	—	—
Full-Scale score[c]	.81	.66	.85	.78	.79	.80	.73	.68	.74	.64	.65	.95	.91

Note. The coefficients of correlation were computed from scaled scores. The average coefficients were computed by transforming each *r* to Fisher's *z* statistic, and reconverting the mean *z* value to the equivalent *r*. The data and table are from Wechsler Adult Intelligence Scale-Revised (Table 16, p. 46) by D. Wechsler, 1981, New York: The Psychological Corporation. Data and table copyright 1989 by The Psychological Corporation and may not be reproduced without permission. All rights reserved.

[a] Verbal score is the sum of scaled scores on the 6 Verbal tests. [b] Performance score is the sum of scaled scores on the 5 Performance tests. [c] Full-Scale score is the sum of scaled scores on all 11 tests. Coefficients with these variables in the main body of the table have been corrected to remove contamination.

differential diagnosis, which is still too often requested by attorneys, insurance companies, and other third-party payers. In regard to the latter, many times during my nearly 40 years as a clinician–teacher providing clinical services to patients in a large university hospital, I have had to address the unreliability of such differential diagnoses offered by me or my psychologist and psychiatrist colleagues. In fact, because the published levels of clinician-to-clinician agreement were so low (*r*s of .20) for the diagnostic categories (e.g., schizophrenia, anxiety neurosis) included in the earlier editions of the *Diagnostic and Statistical Manual of Mental Disorders,* 1st and 2nd editions (*DSM*–I and *DSM*–II; American Psychiatric Association, 1952, 1968), my colleagues and I carried out a federally funded research program (detailed in Matarazzo & Wiens, 1972) in an effort to help identify noncontent parameters of the clinical interview that might help improve such levels of agreement across clinicians and thus help us better serve our patients. This research, begun in 1954, and my first literature review (Matarazzo, 1965), a decade later, of the research of other investigators, plus our own, continued to show that differential diagnoses such as depression, hysteria, and schizophrenia possessed little or no interclinician reliability.

My subsequent review (Matarazzo, 1978) of research during the 1960s and 1970s by investigators in such places as St. Louis, Boston, New York City, and Iowa City indicated a considerable improvement. In fact, when I published my last review of this literature (Matarazzo, 1983b), after the introduction of the *Diagnostic and Statistical Manual of Mental Disorders,* 3rd edition (*DSM*-III; American Psychiatric Association, 1980), the levels of agreement in such differential diagnoses across two independent clinicians had improved materially. Specifically, for many of the discrete diagnostic categories in current use, the levels of interclinician agreement (i.e., *r*s above .80 and .90) now were being reported to be as high as the test–retest reliabilities of the WAIS–R and other well-standardized, objective tests.

What I am discussing in the present section is the reliability (and thus the potential validity) of a discrete, one- or two-word differential diagnosis (e.g., manic depression, obsessive–compulsive disorder, and schizophrenia). That is, because no such body of research has yet been published, I am not discussing the clinician-to-clinician reliability and thus the validity of the personal, social, medical, and psychological portrait of the individual that is typically contained in the comprehensive 10–20-page psychological or neuropsychological assessment of a patient involved in the increasing number of cases also being adjudicated in our nation's courtrooms; four published examples of which I cited above. In my three literature reviews (J. D. Matarazzo, 1965, 1978, 1983b) and related writings (Matarazzo, 1985, 1986), I have tried to present a reasonably accurate picture of the then-current stage of research advances in clinician-to-clinician reliabilities of such single diagnoses, including an informed balance of reliability studies that have reported both negative and positive findings as well as the cogent criticisms of differential diagnosis published by Meehl (1954, 1956, 1957) and Ziskin, to whom I refer later. More recent studies report improvement in reliability from the earlier levels (Grove, 1987), although this is not true for 100% of the publications on the reliability of differential diagnosis between my last review (J. D. Matarazzo, 1983b) and today.

Balanced Literature Review Versus Partisan Scholarship

Unfortunately, a widely publicized review of this same body of literature reached exactly the opposite conclusion. In a three-volume book, *Coping With Psychiatric and Psychological Testimony* (Ziskin & Faust, 1988) and in a summarizing article in *Science* (Faust & Ziskin, 1988a), the authors have attempted to discredit totally the competence of psychologists or psychiatrists to offer a reliable, let alone a valid, psychiatric or psychological differential diagnosis. In the preface to their three-volume book, also published by Ziskin and written almost exclusively for attorneys for use in cross examinations of psychologists and psychiatrists, Ziskin and Faust (1988) accurately, candidly, and commendably have stated that

> The book consists almost entirely of literature which negates the expertise of mental health professionals. There is literature not contained in this book that is supportive of forensic psychiatry and psychology . . . .The reason we exclude supportive literature is *not* so that readers will think it does not exist. As noted, it may or does exist, however, although perhaps of academic interest, we view such supportive evidence as largely irrelevant from a legal context. (p. xvii)

In the three previous editions of this book, which he authored alone, Ziskin included in the preface to each a comparably candid and commendable admission regarding the *lack* of balance, evenhandedness, and scholarly thoroughness of his reviews of the literature.

In a glaring omission from the usual canons of scholarly writing, no such admission regarding the deliberate one-sidedness of the literature review was included by Faust and Ziskin (1988a) in their article in *Science*. Because *Science* is a prestigious journal read by thousands of U.S. scientists and scholars, publication of a lead review article in this journal accords its contents much more credence than would be the case with publication of their companion three-volume book or, for that matter, would be the case if this purported literature review had been published in a journal with a less prestigious reputation in the scholarly and scientific community. As I have experienced, and as Brodsky (1989, p. 261) has independently affirmed, attorneys are already citing this *Science* article in attempts to discredit expert witnesses.

As I have already indicated, in the three literature reviews I published (J. D. Matarazzo, 1965, 1978, 1983b), I tried to include a balance of articles reporting both negative and positive findings regarding the reliability of psychiatric and psychological differential diagnosis. In fact, a number of the negative studies cited in the Ziskin and Faust (1988) book were ones I thoroughly discussed years earlier (Matarazzo, 1965, 1978). What is disappointing in the recent article by Faust and Ziskin (1988a) is that the increasing numbers of more positive studies, many of which were also reviewed in detail by Matarazzo (1983b), were totally omitted in their one-sided review of the literature through 1988 published in *Science*. Brodsky (1989) has sharply drawn attention to this lack of balanced scholarship in the Faust and Ziskin (1988a) article by deliberately entitling his rejoinder "Advocacy in the Guise of Scientific Objectivity: An Examination of Faust and Ziskin."

Throughout the present article I have been critical of my own work and that of other clinical psychologists and clinical neuropsychologists in regard to the bases for some of our conclusions, especially as they relate to the evidence for or against impairment in individuals with suspected brain injuries. However, because of the Faust and Ziskin (1988a) article in *Science*, in the present section I am deliberately defending what we are doing in the field of differential diagnosis relative to the types of other mental health disorders listed in *DSM*–III and its revision, *DSM*–III–R. Specifically, it is my position that, after years of unacceptably low levels of agreement, the test–retest reliability of clinician-to-clinician diagnosis for a number of disorders has improved considerably during the past decade. Although more improvement is necessary and current trends indicate that this improvement will continue over the next decade, research published to date indicates that the levels of reliability now achieved demonstrate moderate to good levels of confidence in many such diagnostic judgments (Grove, 1987; Matarazzo, 1983b).

I must add, lest I too be guilty of one-sided scholarship, that I agree with some of the harsh opinions that Ziskin and Faust (1988; Faust & Ziskin, 1988a, 1989) included in their criticism of the feeble scientific scaffolding currently available to buttress some of the opinions offered by psychologists and psychiatrists who testify in the courtroom. As one such example I cite their three-volume book, which includes a good discussion of a number of studies (albeit with too much emphasis on the earliest ones I had reviewed years ago) that did show poor interclinician agreement. A second example is their detailed discussion of errors they found in the lengthy written psychological assessment reports and subsequent courtroom testimony offered by a few representative psychologists whose work the authors critiqued in an appendix to the third volume of their three-volume work (Ziskin & Faust, 1988). Also, in what I otherwise have described as an unbalanced scholarly review of the published literature, Faust and Ziskin (1988a) and Ziskin and Faust (1988) did cite some other authors who are as critical as they are of the current status of psychiatric and psychological diagnosis. In fact, some of these cited individuals offering such criticisms (e.g., Robins, 1985), following the canons of good science and scholarship, are individuals who themselves have spent much of their professional lives both identifying the need for better reliability and helping make such diagnostic judgments more reliable.

Ziskin and Faust (1988) and Faust and Ziskin (1988a) cited other scholars who are dubious that, for example, *DSM*–III-based diagnoses are reliable. Kutchins and Kirk (1986) offered two criticisms of the *DSM*–III classification system that seem especially cogent to me. The first one deals with the inconsistency across different publication outlets with which the results obtained and the methodology used in the development of *DSM*–III were described. Specifically, the values of the clinician-to-clinician coefficients of reliability of diagnosis that were published in a 1979 issue of the *American Journal of Psychiatry* by Spitzer, Forman, and Nee (1979) were, without adequate explanation, different from the values of the comparable reliability coefficients subsequently published in the 1980 *DSM*–III. Furthermore, the 1979 article indicated that each clinician in the reliability study "was expected to participate with another clinician in at least two reliability evaluations" (Spitzer et al., 1979, p. 815). However, this identical component of the methodology used was reported as "Each of these clinicians was asked to participate in at least four reliability evaluations with another clinician" in both the 1980 *DSM*–III (p. 46) and in the subsequent *Archives of General Psychiatry* article by Hyler, Williams, & Spitzer (1982, p. 1275).

Kutchins and Kirk (1986), appropriately, have also pointed out that the number of clinicians participating in the development of *DSM*-III was inadequately explained, inasmuch as 365 clinicians were cited in one publication and only 274 clinicians in another, with no explanation for the seeming 25% attrition. Furthermore, the number of patients evaluated by these 274 clinicians was reported as 281 in Spitzer et al. (1979) and 339 in the 1980 *DSM*–III. These critics also pointed out that the actual percentages in which the two clinicians conducted the reliability interviews, jointly or separately, was reported differently in Spitzer et al. and in the *DSM*–III. In their own judgment of this first family of criticisms, Kutchins and Kirk wrote "Such discrepancies as these (and there may be others) may not be serious, but they illustrate some of the difficulty in understanding exactly what was done in the (DSM III) field trials" (p. 5).

A second cogent criticism offered by Kutchins and Kirk (1986) deals with the low magnitudes of the clinician-to-clinician levels of agreement in that part of the *DSM*–III study concerning diagnoses of disorders in children. Kuchins and Kirk pointed out that the magnitudes of kappa reported in the *DSM*–III for agreement between two clinicians for the diagnoses of disorders in children and adolescents are at once lower than those for adult disorders and are unacceptably low in their magnitudes. They added that, even for the disorders of adults, the magnitudes of kappa reported show far from perfect clinician-to-clinician agreement and the interpretation of the magnitudes has been inconsistent, varying, for example, from "reliability for most classes [of *DSM*–III] was quite good" to the kappa values for personality disorders were "quite low" and were "only fair" for the disorders of children.

Although I agree with the gist of these two criticisms, I do not agree with other criticisms of Kutchins and Kirk (1986). Part of my disagreement involves seeing the bottle of wine as half empty versus seeing it as half full. I agree that Kutchins and Kirk are correct in the somewhat pessimistic tone in which they cast their perception that the magnitudes of kappa characterizing the reliability of diagnoses for almost all classifications published to date fail to reach 1.00; and furthermore, I agree that the values of clinician-to-clinician agreement for some disorders indicate that the reliabilities of such diagnoses are not better than could be obtained by chance.

However, from my perspective, having followed this area of research during the past three decades (Matarazzo, 1965, 1978, 1983b), the degree of agreement published during the past decade (more examples of which are found in the *Archives of General Psychiatry* and the *American Journal of Psychiatry*)—whereas not yet meeting acceptable canons of science for each and every one of the extant diagnostic categories, as discussed in more detail in Matarazzo (1983a) and in Grove (1987)—do meet such concerns for a relatively large number of diagnostic categories. In addition, the confirmatory studies independently published by other investigators since my 1983 literature review have increased in quantity and in the numbers of diagnostic categories that show improvement in reliability.

Kutchins and Kirk (1986) offered still other criticisms that I do not find compelling. First, their concern that the computation of kappa in the studies reported in the *DSM*–III is unclear is a criticism with less effect when one considers, first, that its computation is straightforward (see the example in Matarazzo, 1965 or 1978) and, second, that dozens of studies using kappa have been published independently by investigators other than Spitzer and the team that developed *DSM*–III. Equally unconvincing because the problem has been amply addressed by other investigators is Kutchin and Kirk's suggestion that investigators need to supplement the published values of kappa with concurrent publication in the same tables of other statistical properties that impact the interpretation of kappa. These include (a) sensitivity (the proportion of time each clinician in the study of test–retest reliability made a *positive* diagnosis when a disorder was present), (b) specificity (the proportion of time each clinician made a *negative* diagnosis when a disorder was absent), and (c) base rate (the prevalence of the disorder in the sample of patients being studied). I find this 1986 criticism by Kutchins and Kirk uncompelling because dozens of published studies have, in fact, included values for these additional variables (Matarazzo, 1983b). A few earlier studies and some more recent studies have addressed still other relevant variables needed for interpreting reliability levels such as *Cronbach's alpha,* a measure of the internal consistency of the diagnoses (e.g., Widiger, Trull, Hurt, Clarkin, & Frances, 1987), and *bias,* the extent to which errors in diagnosis tend to be made more in one direction than the other and thus lead to false estimates of prevalence (e.g., Robins, 1985). In fact, Widiger, Hurt, Frances, Clarkin, and Gilmore (1984) offered a sophisticated analysis of a number of these statistical measures with which to improve on the inappropriate use of kappa in studies on the reliability of diagnosis.

My impression is that whereas 30 years ago almost *all* of the published studies relating to the degree of agreement on differential diagnosis of the disorders then listed in the *DSM* produced results that showed poor clinician-to-clinician reliability, my reading suggests that about 50% of the studies published in the past decade report good to very good magnitudes of reliability. Also, the trend line suggests that the percentage of studies reporting acceptable levels of interclinician reliability will increase even further during the next decade.

Informed Consumer Acceptance as Interim Validation

Research that demonstrates the validity (e.g., treatment studies that show that antipsychotic medication is efficacious in treating schizophrenia) of such single- or two-word differential diagnoses, although available, is considerably more sparse for mental disorders (see Feighner & Herbstein, 1987; Matarazzo, 1978), although it is considerably more than adequate for mental retardation and the various gradations of intellectual ability (see Matarazzo, 1972, chapters 5, 6, 7, and 12). Therefore, in regard to the critical issue that the validity of *DSM*–III-type differential diagnoses has not been adequately established, Faust and Ziskin (1989) and I (Matarazzo, 1978) are in agreement. However, in excerpting my views on this issue, Faust and Ziskin (1988b, p. 1144; 1989, pp. 33–34) have selectively taken passages from my writings out of context. They correctly quote my belief that currently there is no body of research that indicates that psychological assessment across the whole domain is valid or is other than clinical art. However, they neglect to add that I include in those same passages (Matarazzo, 1985, pp. 247–248; 1986, p. 20) the equally relevant opinion that in this regard psychology is little different from engineering, medicine, or other professions. That is, professions in which practitioners' (artisans') work products are judged by society to be valid (usable) for many services, despite the absence of the necessary research, primarily on the basis that common experience (of legislators, professional peers, patients, clients, and others) suggests some utility from their services. In those published passages I add that the acceptance by these varied constituencies of a qualified practitioner's work product as probably being valid comes only after a professional engineer, physician, psychologist, and other practitioner has (a) first met a set of educational requirements, (b) passed an examination and has become licensed or comparably accredited by the state, (c) had an in-depth review of samples of his or her professional work by members of a specialty board of professional peers, (d) routinely shared and thus has had reviewed some of his or her clinical work products by peers who also are professionally involved with this client or patient, and (e) had those who have paid for and received such services conclude that the services were beneficial.

Meehl (1973), writing as an experienced practitioner and acknowledging that all the formal diagnostic classifications then extant were other than perfectly reliable, expressed a similar opinion when he wrote,

> It is not true that formal nosological diagnosis in psychiatry is as unreliable as the usual statements suggest. If we confine ourselves to major diagnostic categories (e.g., schizophrenia versus nonschizophrenia, organic brain syndrome versus functional disorder, and the like), if we require adequate clinical exposure to the patient (why would anyone in his right mind conduct a study of diagnostic rubrics based upon brief outpatient contact?), and if we study well-trained clinicians who take the diagnostic process seriously, then it is not clear that interclinician diagnostic agreement in psychiatry is worse than in other branches of medicine. (A colleague responds with "That's true, but medical diagnoses are completely unreliable also." I am curious what leads this colleague, given his "official" classroom beliefs, to consult a physician when he is ill? Presumably such an enterprise is pointless, and taking your sick child to a pediatrician is wasted time and money. Do any of my readers *really* believe this?) (p. 273)

That issue having been addressed, my purpose here was not to defend again, as I tried to do in 1978 and 1983, either the reliability or the validity of the diagnoses that are included in *DSM*–III or in related classification systems. Rather, it was to point out that because a beginning scientific scaffolding currently exists, reliable and valid psychological assessment, especially of cognitive functioning in brain injury, *is* being carried out. In the earlier sections of this article I have been critical of my own work and that of other clinicians involved in such assessment; however, even without adequate validation that research of the type I predict will be done before long, my experience in the courtroom has persuaded me that when such assessment is done well, it is patently obvious to all involved (i.e., juries, judges, and the attorneys for *both* the plaintiff and defense) that what such a psychologist–expert-witness concluded was valid (true) within the reasonable degree of certainty required in such litigation.

I will close by citing summaries of two examples of valid psychological assessment (portrait) findings. The first is that of a 21-year-old college history major with high school SATs in the 98th percentile, who had already been inducted into Phi Beta Kappa and who, following a serious automobile injury to her brain, now earned a WAIS–R FS IQ of 74 (3rd percentile), with comparably low scores on other test batteries. An expert-witness psychologist examined her, and in her 16-page report she interpreted these WAIS–R and other test–suggested deficits in the context of the patient's earlier, well-researched and described life history and in the clear-cut, objective postinjury findings recorded in medical and hospital records. The second example is of a patient who alleged that his recent poor memory and related loss of cognitive functioning was due to exposure to neurotoxins in the workplace; however, all medical and related objective findings were negative, and a review of his school and military records by the psychologist retained by the defendant's attorney revealed that he had always functioned cognitively in the lowest 25th percentile of his peers. Opposing attorneys accepted these psychological assessments, and both suits were settled out of court.

Research published much earlier showed that the types of one- or two-word differential diagnoses, characterizations, and predictions then extant were judged to be lacking in validity (Meehl, 1954, 1956, 1957). Reviews of more current studies (Dawes, Faust, & Meehl, 1989), including an excellent recent update of the use of one's head instead of formulas (Kleinmuntz, 1990) reaffirm that conclusion. However, I know of no research (see Korchin, 1976, pp. 258–260) to date regarding the validity of the psychological portraits offered as expert opinion of the type involved in the two aforementioned cases: that is, assessment done by a well-trained clinician familiar with the types of literature I discussed earlier in this article and one who takes such diagnosis as seriously as is suggested by Meehl (1973, pp. 272–281) and here by myself. It is my hope that empirical research on such state-of-the-art psychological assessment will soon be undertaken. When it is, I firmly believe that research will reveal that acceptable levels of validity do now exist for these modern comprehensive psychological assessments and that it will serve as the requisite empirical basis for the consensual agreement regarding the validity of such expert opinions currently being reached by the attorneys for both the plaintiff and defendant for that subset of cases that I know first hand are being settled without going to court.

Earlier in this article, I described clinically relevant psychometric issues and related literature with which a subset of psychologists are not familiar, and in part for that reason, the cases do go to court for adjudication. Inasmuch as increasing numbers of attorneys are becoming familiar with the psychometric properties of psychological tools, it is incumbent upon psychologist–clinicians to be at least as familiar as are they with the strengths and weaknesses of the instruments currently used in psychological assessment. The result will be to increase further the numbers of psychological assessment portraits and characterizations that *both* attorneys agree seem valid. Thus the evidence for such validity will not need to be argued and litigated in the courtroom, but instead will continue to be improved and reported as before in our scientific journals.

REFERENCES

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders* (1st ed.). Washington, DC: Author.

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: Author.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

Amrine, M. (Ed.). (1965). Special issue: Testing and public policy. *American Psychologist, 20,* 857–993.

Bowman, M. L. (1989). Testing individual differences in Ancient China. *American Psychologist, 44,* 576–578.

Brodsky, S. L. (1989). Advocacy in the guise of scientific advocacy: An examination of Faust and Ziskin. *Computers in Human Behavior, 5,* 261–264.

Cohen, R. J., Montague, P., Nathanson, L. S., & Swerdlik, M. E. (1988). *Psychological testing: An introduction to tests and measurement.* Mountain View, CA: Mayfield.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674.

Doyle, K. O., Jr. (1974). Theory and practice of ability testing in Ancient Greece. *Journal of the History of the Behavioral Sciences, 10,* 202–212.

DuBois, P. H. (1970). *A history of psychological testing.* Boston: Allyn & Bacon.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86,* 335–337.

Elliott, R. (1987). *Litigating intelligence: IQ tests, special education, and social science in the courtroom.* Dover, MA: Auburn House.

Faust, D., & Ziskin, J. (1988a). The expert witness in psychology and psychiatry. *Science, 241,* 31–35.

Faust, D., & Ziskin, J. (1988b). Response to Fowler and Matarazzo. *Science,* 1143–1144.

Faust, D., & Ziskin, J. (1989). Computer-assisted psychological evidence as legal evidence: Some day my prints will come. *Computers in Human Behavior, 5,* 23–36.

Feighner, J. P., & Herbstein, J. (1987). Diagnostic reliability. In C. G. Last & M. Hersen (Eds.), *Issues in diagnostic research* (pp. 121–140). New York: Plenum.

Goldstein, G. (1984). Comprehensive neuropsychological assessment batteries. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 181–210). New York: Pergamon Press.

Goldstein, G., & Shelly, C. H. (1971). Field dependence and cognitive, perceptual and motor skills in alcoholics. *Quarterly Journal of Studies on Alcohol, 32,* 29–40.

Goldstein, G., & Shelly, C. H. (1972). Statistical and normative studies of the Halstead Neuropsychological Test Battery relevant to a neuropsychiatric hospital setting. *Perceptual and Motor Skills, 34,* 603–620.

Grove, W. M. (1987). The reliability of psychiatric diagnosis. In C. G. Last & M. Hersen (Eds.), *Issues in diagnostic research* (pp. 99–119). New York: Plenum.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

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

Heilbronner, R. L., Buck, P., & Adams, R. L. (1989). Factor analysis of verbal and nonverbal clinical memory tests. *Journal of Clinical Neuropsychology, 4,* 299–309.

Hyler, S. E., Williams, J. B. W., & Spitzer, R. L. (1982). Reliability in the *DSM*-III field trials: Interview and case summary. *Archives of General Psychiatry, 39,* 1275.

Kleinmuntz, B. (1990). Why we still use our heads instead of the formulas: Toward an integrative approach. *Psychological Bulletin, 107,* 296–310.

Korchin, S. J. (1976). *Modern clinical psychology.* New York: Basic Books.

Kupke, T., & Lewis, R. (1989). Relative influence of subject variables and neurological parameters on neuropsychological performance of adult seizure patients. *Archives of Clinical Neuropsychology, 4,* 351–363.

Kutchins, H., & Kirk, S. A. (1986, Winter). The reliability of *DSM*-III: A critical review. *Social Work Research and Abstracts, 22,* 3–12.

Leckliter, I. N., & Matarazzo, J. D. (1986). A literature review of factor analytic studies of the WAIS–R. *Journal of Clinical Psychology, 42,* 332–342.

Leckliter, I. N., & Matarazzo, J. D. (1989). The influence of age, education, IQ, gender, and alcohol abuse on Halstead-Reitan Neuropsychological Test Battery performance. *Journal of Clinical Psychology, 45,* 484–511.

Lippmann, W. (1922). *New Republic, 32,* 9–10, 213–215, 246–248, 275–277, 297–298, 328–330.

Matarazzo, J. D. (1965). The interview. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 403–450). New York: McGraw-Hill.

Matarazzo, J. D. (1972). *Wechsler's measurement and appraisal of adult intelligence: Fifth and enlarged edition.* New York: Oxford University Press.

Matarazzo, J. D. (1978). The interview: Its reliability and validity in psychiatric diagnosis. In B. B. Wolman (Ed.), *Clinical diagnosis of mental disorders: A handbook* (pp. 47–96). New York: Plenum.

Matarazzo, J. D. (1983a). Computerized psychological testing (Editorial). *Science, 221,* 323.

Matarazzo, J. D. (1983b). The reliability of psychiatric and psychological diagnosis. *Clinical Psychology Review, 3,* 103–145.

Matarazzo, J. D. (1985). Clinical psychological test interpretations by computer: Hardware outpaces software. *Computers in Human Behavior, 1,* 235–253.

Matarazzo, J. D. (1986). Computerized clinical psychological test interpretations: Unvalidated plus all mean and no sigma. *American Psychologist, 41,* 14–24, 94–96.

Matarazzo, J. D., Carmody, T. P., & Jacobs, L. D. (1980). Test–retest reliability and stability of the WAIS: A literature review with implications for clinical practice. *Journal of Clinical Neuropsychology, 2,* 89–105.

Matarazzo, J. D., Daniel, M. H., Prifitera, A., & Herman, D. O. (1988). Intersubtest scatter in the WAIS–R standardization sample. *Journal of Clinical Psychology, 44,* 940–950.

Matarazzo, J. D., & Herman, D. O. (1984a). Relationship of education and IQ in the WAIS–R standardization sample. *Journal of Consulting and Clinical Psychology, 52,* 631–634.

Matarazzo, J. D., & Herman, D. O. (1984b). Base-rate data for the WAIS–R: Test–retest stability and VIQ–PIQ differences. *Journal of Clinical Neuropsychology, 6,* 351–366.

Matarazzo, J. D., & Herman, D. O. (1985). Clinical uses of the WAIS–R: Base rates of differences between VIQ and PIQ in the WAIS–R standardization sample. In B. B. Wolman (Ed.), *Handbook of Intelligence: Theories, measurements and applications* (pp. 899–932). New York: Wiley.

Matarazzo, J. D., Matarazzo, R. G., Wiens, A. N., Gallo, A. E., & Klonoff, H. (1976). Retest reliability of the Halstead Impairment Index in a normal, a schizophrenic, and two samples of organic patients. *Journal of Clinical Psychology, 32,* 338–349.

Matarazzo, J. D., & Prifitera, A. (1989). Subtest scatter and premorbid intelligence: Lessons from the WAIS–R standardization sample. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1,* 186–191.

Matarazzo, J. D., & Wiens, A. N. (1972). *The interview: Research on its anatomy and structure.* Chicago: Aldine-Atherton.

Matarazzo, J. D., Wiens, A. N., Matarazzo, R. G., & Goldstein, S. G. (1974). Psychometric and clinical test–retest reliability of the Halstead Impairment Index in a sample of healthy, young, normal men. *Journal of Nervous & Mental Disease, 158,* 37–49.

Matarazzo, R. G., Matarazzo, J. D., Gallo, A. E. Jr., & Wiens, A. N. (1979). IQ and neuropsychological changes following carotid endarterectomy. *Journal of Clinical Neuropsychology, 1,* 97–116.

Matarazzo, R. G., Wiens, A. N., Matarazzo, J. D., & Manaugh, T. S. (1973). Test–retest reliability of the WAIS in a normal population. *Journal of Clinical Psychology, 29,* 194–197.

Meehl, P. E. (1954). *Clinical vs. statistical prediction.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1956). Wanted—A good cookbook. *American Psychologist, 11,* 263–272.

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology, 4,* 268–273.

Meehl, P. E. (1973). Why I do not attend case conferences. In P. E. Meehl (Ed.), *Psychodiagnosis: Selected papers* (pp. 225–302). Minneapolis: University of Minnesota Press.

Rapaport, D., Gill, M., & Schafer, R. (1945). *Diagnostic psychological testing.* Chicago: Yearbook Publishers.

Robins, L. N. (1985). Epidemiology: Reflections on testing the validity of psychiatric interviews. *Archives of General Psychiatry, 42,* 918–924.

Royce, J. R., Yeudall, L. T., & Bock, C. (1976). Factor analytic studies of human brain damage: I. First and second-order factors and their brain correlations. *Multivariate Behavioral Research, 4,* 381–418.

Silverstein, A. B. (1985). Cluster analysis of the Wechsler Adult Intelligence Scale–Revised. *Journal of Clinical Psychology, 41,* 98–100.

Sokal, M. M. (Ed.). (1987). *Psychological testing and American society: 1890–1930.* New Brunswick, NJ: Rutgers University Press.

Spitzer, R. L., Forman, J. B. W., & Nee, J. (1979). *DSM*–III field trials: I. Initial interrater diagnostic reliability. *American Journal of Psychiatry, 136,* 815–817.

Swiercinsky, D. P. (1979). Factorial pattern description and comparison of functional abilities in neuropsychological assessment. *Perceptual and Motor Skills, 48,* 231–241.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale–Revised.* New York: Psychological Corporation.

Wechsler, D. (1987). *Wechsler Memory Scale–Revised.* San Antonio, TX: Psychological Corporation.

Widiger, T. A., Hurt, S., Frances, A., Clarkin, J., & Gilmore, M. (1984). Diagnostic efficiency and *DSM*–III. *Archives of General Psychiatry, 41,* 1005–1012.

Widiger, T. A., Trull, T. J., Hurt, S. W., Clarkin, J., & Frances, A. (1987). *DSM*–III Personality disorders. *Archives of General Psychiatry, 44,* 557–563.

Ziskin, J., & Faust, D. (1988). *Coping with psychiatric and psychological testimony* (Vols. 1–3, 4th ed.). Marina Del Ray, CA: Law & Psychology Press.