DISCUSSION 6

Prior to beginning work on this discussion, read Chapter 12 in the textbook and the required articles for this week. For this discussion you will take on the role of a psychologist assigned a case in which the client has a legal concern. For your initial post, select one of the three forensic case scenarios below and follow the instructions.

Forensic Scenario One: Mr. W (Attempting to Obtain Legal Guardianship Over an Elderly Parent): Attorney Mr. X referred Mr. W for an evaluation of his decision-making capacity. Mr. W's children do not agree with the findings from a prior evaluation and have requested a second opinion. Review the PSY640 Week Six Clinical Neuropsychological Report for Mr. W (Links to an external site.), and begin your post with a one-paragraph summary of the test data you deem most significant. Utilize assigned readings and any additional scholarly and/or peer-reviewed sources needed to develop a list of additional assessment instruments and evaluation procedures to administer to the client. Justify your assessment choices by providing an evaluation of the ethical and professional practice standards and an analysis of the reliability and validity of the instruments

CLINICAL NEUROPSYCHOLOGY REPORT

Patient's Name: Mr. W **Date of Evaluation: 10/10/2014**
Date of Birth: 10/02/24 **Age: 90** **Handedness: Right**
Education: 6 years **Occupation: City worker (retired)**
Current Medications: Donepezil 5 mg/day, Simvastatin 40 mg/day, Levothyroxin 1.25 mg/day, Losartan 50 mg/day, Warfarin 3 mg/day, Advair Inhaler, Ventolin Inhaler, Alendroate Sodium 35 mg/week, Vitamins B12 and D3

Evaluation Completed by: Dr. K., Ph.D.
Evaluation Time: One hour diagnostic interview (90791); One hour test administration, scoring, interpretation and report (96118 x 3)

REASON FOR REFERRAL: Attorney Mr. X referred Mr. W for an evaluation of his decision-making capacity.

HISTORY OF CURRENT SYMPTOMS: The symptom description and history were obtained from an interview with Mr. W, his sister, and his cousin. Mr. W stated he was seen by a physician in Michigan last year at his son's urging and was diagnosed with "dementia." Subsequently, according to the patient, his son reportedly took control of his finances, has withdrawn approximately $28,000 from the patient's account, and has sold the patient's coin collection. Mr. W does not feel the diagnosis of dementia is correct and would like to resume control over his financial matters.

Reportedly, the incident that initiated the diagnosis of dementia occurred in 2011 when Mr. W was living with his son Anthony. He stated he saw the silhouette of a person walking in another room in the house and believed it was the "Boogie Man." Several days later, he had what appeared to be a syncopal episode ("I blacked out") and fell while walking out to the garage. He stated he felt someone "pounding my head and pulling me down the stairs," and he believed this was also the "Boogie Man". He was reportedly taken to the ER and released; however, after this incident the patient stated his sons became concerned with his thinking, and this eventually led to an evaluation with a physician and a diagnosis of dementia.

Mr. W denied any other instances or auditory or visual hallucinations beyond those described above. He was living in A State (initially with his family and then on his own), but in 20XX, moved to Another State to live with his sister and brother-in-law. According to his sister and his cousin, the patient has not demonstrated any problems with memory or other areas of thinking. He stopped driving two years ago at the insistence of his son, but he remains independent in other activities of daily living, including managing his own medications, self-care, and occasional household chores. He also enjoys playing cards and playing electronic poker, and there has been no reported decline in his ability in these areas.

Summary of Previous Investigations and Findings: No previous neuropsychological evaluations.

PAST MEDICAL, NEUROLOGICAL, PSYCHIATRIC, SUBSTANCE USE HISTORY: *(Inclusive review of symptoms and disorders; only positive features listed)* Hypertension, hypercholesterolemia, hypothyroidism, COPD, asthma, myocardial infarction in the past (exact date unknown), and osteoporosis. The patient denied any neurological or psychiatric history beyond that described above. He does not drink alcohol and quit smoking in 1940. He has no history of recreational drug use.

BIRTH, DEVELOPMENTAL, OCCUPATIONAL HISTORY: *(Review of perinatal factors, early childhood development and milestones, academic history and achievement, employment)* No reported delays in reaching developmental milestones. The patient stated he completed 6 years of formal education and worked for the city in the sewer division for many years.

FAMILY HISTORY: *(First degree relatives; only pertinent features reported)* The patient's mother reportedly died of a stroke at age 57, and the patient's father died in an accident when the patient was 14. The patient has one full brother, age 81, who is reportedly in good health, and one half-brother with whom

he does not have regular contact. The patient has five children (three sons and two daughters), but he and his wife did not live together consistently at the time the children were born, so he stated he is not sure he is the biological father of his three oldest children. He reported he currently has no ongoing contact with any of his children.

PSYCHOSOCIAL HISTORY AND CURRENT ADAPTATION: *(Current living situation, social relationships, activities of daily living)* The patient lived in A State most of his life, but moved to Another State to be closer to his children about a year ago. He was living with his son and then Another Son until 20XX when he moved into an independent apartment. He lived alone for one year before he moved to Another State to live with his sister and brother-in-law due to his ongoing conflicts with his son regarding financial issues.

CURRENT EXAMINATION: *Review of records; Clinical Interview; Cognitive Assessment: Wechsler Test of Adult Reading (WTAR); Wechsler Adult Intelligence Scale-IV (WAIS-IV) (partial); Attention Tests: WAIS-IV Digit Span, Trail Making Tests, RBANS Coding, RBANS Semantic Fluency; Language Tests: RBANS Naming Test; Visuospatial Tests: RBANS Figure Copy and Line Orientation, Target cancellation; Learning/Memory Tests: RBANS Word List, Story and Figure recall; Reasoning/Abstraction: WAIS-IV Similarities*

BEHAVIORAL OBSERVATIONS:
The patient arrived on time for his appointment and was accompanied by his sister and his cousin. He was casually dressed and neatly groomed, and his social interpersonal skills were preserved. He was very pleasant and put forth good effort throughout the evaluation. Thought processes were logical and goal directed, and there was no indication of hallucinations, delusions, or other psychoses. No overt behavioral indications of a mood disturbance were observed, and a full range of affect was demonstrated.

The results of this evaluation are considered reliable and valid for interpretation.

SUMMARY OF FINDINGS:
Based on his educational history (6th grade) and performance on the WTAR (est. FSIQ = 68) the patient's estimated level of premorbid functioning would be within the low-average to borderline range overall. The remainder of the examination was interpreted with the expectation of performance at this level.

The patient was fully oriented with the exception of the city, which he did not know. He was able to give detailed information (e.g., specific dates) of his autobiographical history, and his performance on formal memory testing did not indicate any type of retentive memory disturbance. Although he had slight difficulty encoding new information, there was no loss of information over time.

The patient's speech was fluent with normal articulation, and rate and comprehension of auditory information was intact. No significant impairments were noted in naming, reading, or writing. Visuospatial abilities were an area of relative weakness, but there was no indication of hemispatial neglect or inattention, and object recognition was preserved. It is likely his poor performance on the RBANS Figure Copy and Line Orientation was due to difficulties in higher level visuospatial processing and executive functions. Abstract verbal reasoning was within normal parameters.

Immediate attention span was intact, and he performed within normal limits on most tests of sustained attention. His score on the RBANS coding subtest, which also has a visuospatial and motor component, was the only area that was below expectation.

TESTING SUMMARY:	09/10/2011	Normative data	Current Level*
PREMORBID FUNCTIONING			
WTAR	10/50	SS = 68	Borderline/Low
DEMENTIA SCREENING			
MMSE	25/30	--	Within Normal Limits

ATTENTION			
WAIS-IV Digit Span	5 F, 5 B	ss = 9	Average
RBANS Coding	20/89	ss = 4	Borderline/Low
Trail Making Test Part A	49"	T = 53	Average
Trail Making Test Part B	115"	T = 62	High Average
LANGUAGE			
RBANS Naming	10/10	>75th%	High Average
RBANS Semantic Fluency	16 words/min	ss = 9	Average
VISUOSPATIAL			
RBANS Figure Copy	10/20	ss = 2	Extremely Low
RBANS Line Orientation	4/20	<2nd%	Extremely Low
MEMORY			
RBANS Word List			
Learning Trials	17/40	ss = 6	Low Average
Delayed Recall	0/10	3-9th%	Borderline
Recognition	19/20	26-50th	Average
RBANS Story			
Learning Trials	8/24	ss = 4	Borderline/Low
Delayed Recall	6/12	ss = 8	Average
RBANS Figure Recall	6/20	ss = 6	Low Average
EXECUTIVE FUNCTIONS			
WAIS-IV Similarities	--	ss = 5	Borderline

REPEATABLE BATTERY FOR THE ASSESSMENT OF NEUROPSYCHOLOGICAL STATUS*:		
Index Scores	Mean = 100; std = 15	Current Level
Immediate Memory	SS = 78	Borderline
Visuospatial/Constructions	SS = 53	Extremely Low
Language	SS = 99	Average
Attention	SS = 68	Borderline/Low
Delayed Memory	SS = 90	Average

*80-89 year-old norms used because 90 year-old-norms are not available

SUMMARY AND IMPRESSION:

1. *Neurocognitive Profile:* The profile on testing is one of mild weaknesses in some aspects of complex attention/working memory and executive functions within the context of an overall low average to borderline level of general intellectual functioning. Although his primary visuospatial abilities are intact, he demonstrated a weakness on more complex visuospatial processing, most likely due to the executive aspects of these tasks. He had some difficulty initially encoding lengthy (e.g., story) information, but delayed recall and recognition were generally intact, and there is no indication of a primary retentive memory disturbance. The patient did not endorse any symptoms consistent with a mood disturbance and there was no indication of hallucinations, delusions, or other psychoses observed during the interview and examination.

2. *Diagnostic Formulation:* The profile on testing is consistent with a mild dysfunction in frontal networks. In this case, the differential diagnosis is extensive and includes potential cerebrovascular disease (given his risk factors and history of at least one syncopal episode) and toxic/metabolic abnormalities (e.g., thyroid abnormalities). The etiology of his syncopal episode and confusion is impossible to determine in the absence of medical records from that time, but his hallucinations during that time are consistent with his religious and spiritual beliefs. In addition, there have been no further instances or evidence of hallucinations or other psychoses to suggest this is an ongoing/active problem. Although the possibility can never be fully excluded in this age group, the absence of retentive memory impairment argues

strongly against the likelihood that Alzheimer's disease is the primary, or a significant cause of, his current cognitive symptoms.

RECOMMENDATIONS:

1. Mr. W's cognitive weaknesses are not sufficient to render him incapable of making his own decisions regarding his finances and/or health care, and therefore, guardianship is not appropriate.

2. Mr. W should continue to refrain from operating a motor vehicle or engaging in any potentially dangerous activities (such as the use of heat generating appliances or power tools) due to his visuospatial and attentional weaknesses.

3. Mr. W was encouraged to follow-up with his primary care physician to a) ensure that all treatable causes of cognitive impairment are well-controlled (e.g., thyroid, blood pressure, diabetes, etc.), and b) review and update his medications. He may also want to discuss with his doctor whether a neurological work-up (including some form of brain imaging) would be helpful to further clarify the etiology of his current cognitive symptoms

4. A follow-up evaluation can be conducted in the future if there is evidence of symptom change or progression.

_____, Ph.D., ABPP-CN
Board Certified Neuropsychologist
Licensed Clinical Psychologist

cc: Mr. X, Attorney at Law
Dr. Diaz
Mr. W

Gregory, R.J. (2014). Psychological testing: History, principles, and applications (7th ed.). Boston, MA: Pearson

CHAPTER 12
Legal Issues and the Future of Testing

TOPIC 12A Psychological Testing and the Law

12.1 The Sources and Nature of Law
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec1#ch12lev1sec1)

12.2 Testing in School Systems and the Law
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec2#ch12lev1sec2)

12.3 Disability Assessment and the Law
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec3#ch12lev1sec3)

12.14 Legal Issues in Employment Testing
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec4#ch12lev1sec4)

Case Exhibit 12.1 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec4#ch12box2)
Unwise Testing Practices in Employee Screening
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec4#ch12box2)

12.5 Forensic Applications of Assessment
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec5#ch12lev1sec5)

In the previous chapters we have outlined the myriad of ways in which tests are used in decision making. Furthermore, we have established that psychological testing is not only pervasive, but it is also consequential. Test results matter. Test findings may warrant a passage to privilege. Conversely, test findings may sanction the denial of opportunity. For many reasons, then, it is appropriate to close the book with two special topics that bear upon the potential repercussions of psychological testing. In **Topic 12A (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12#ch12box1)** , Psychological Testing and the Law, we review critical legal issues pertaining to the use of psychological tests. In this topic, we survey the essential laws that regulate the use of tests in a variety of settings—schools, employment situations, medical settings, to name just a few arenas in which the law constrains psychological testing. We also examine several ways that psychologists interface with the legal system in the field of forensic assessment. In **Topic 12B (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec5#ch12box3)** , Computerized Assessment and the Future of Testing, contemporary applications of the computer in psychological assessment are surveyed, and then the professional and social issues raised by this practice are discussed. The book closes with thoughts on the future of testing—which will be forged in large measure by increasingly sophisticated applications of computer technology but also greatly affected by legal standards.

12.1 THE SOURCES AND NATURE OF LAW

The law establishes a number of guidelines that define the permissible scope and applications of psychological testing. However, before investigating the key legal guidelines that impact testing, it will be helpful to understand the sources and nature of law. Broadly speaking, there are three sources of law: constitutional provisions, legislative edicts, and judicial opinions. We examine each briefly.

Constitutional Sources of Law

The United States has a constitutional form of government, meaning that the U.S. Constitution is the final authority for all legal matters in the country. All other forms of law must be consistent with this seminal document. Thus, the Constitution places limits on legislative actions and judicial activity. The United States is also a federation of states, which means that each state retains its own government and system of laws, while ceding some powers to the central government. For example, the power to regulate interstate commerce and the responsibility to provide for the national defense both reside with the federal government. Each state has its own constitution as well, which is another source of laws that affects citizens living in a state. Of course, state constitutions cannot contradict the U.S. Constitution and, in most cases, they are highly similar to the federal document.

Three provisions of the U.S. Constitution potentially bear upon the practice of psychological testing: the Fifth, Sixth, and Fourteenth Amendments to the Constitution (**Melton et al., 1998** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)**). The Fifth Amendment provides a privilege against self-incrimination, which impacts the nature of psychological assessment in forensic evaluations. For example, as discussed previously, a forensic practitioner might be asked by the court to evaluate an alleged offender for competency to stand trial. In many states, self-incriminating disclosures made during an evaluation of competency to stand trial cannot be used to determine guilt (i.e., they are inadmissible as evidence during trial).

The Sixth Amendment states that every person accused of a crime has the right to counsel (i.e., the right to a lawyer). This is understood to mean both the presence of counsel during legal proceedings and also the right to effective assistance from counsel. Does this mean that counsel must be present during a pretrial assessment, such as a court-ordered evaluation for competency to stand trial? This will depend upon the state and jurisdiction in which the proceedings occur. Although most courts have held that the defendant does not have a right to the presence of counsel during pretrial psychological evaluations, a minority of courts have held that the Sixth Amendment guarantee does apply to such pretrial assessments (**Melton et al., 1998** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)**). In these jurisdictions, the defendant's lawyer can be present during any psychological testing or evaluation. This raises difficult questions as to the validity of assessments undertaken in the presence of a third party. For example, what if the client asks his or her lawyer for advice on how to answer certain questions? Surely, this is not standard protocol in psychological assessment and might drastically affect the validity of the results. Fortunately, most courts favor alternative methods for protecting the rights of defendants during pretrial evaluations, such as tape-recording the session, having a defense psychologist observe the evaluation, or providing for an independent evaluation.

The Fourteenth Amendment provides that no state shall deprive any U.S. citizen of life, liberty, or property without "due process of law." The amendment also specifies "equal protection of the laws." The relevant section reads:

> No State shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any State deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws.

It is mainly the "due process" feature of this amendment that has impacted psychological practice. This influence is limited largely to forensic practitioners who deal with competency to stand trial, civil and criminal commitment, or the right to refuse treatment. For example, psychologists who are involved in the civil commitment of an individual who needs treatment typically must show—as a direct consequence of the due process clause of the Fourteenth Amendment—that several stringent criteria are fulfilled:

- The individual must be reliably diagnosed as suffering from severe mental illness;
- In the absence of treatment, the prognosis for the individual is major distress;
- The individual is incompetent; that is, the illness substantially impairs the person's ability to understand or communicate about the possibility of treatment;
- Treatment is available;
- The risk–benefit ratio of treatment is such that a reasonable person would consent to it. (**Melton et al., 1998** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128), p. 310)

Whether these conditions are met would be determined at a commitment hearing during which the individual would have full procedural rights such as the presence of counsel. The psychologist's role would be to offer professional opinions on these guidelines. Of course, the validity of psychological assessment is relevant to these criteria in several ways, including the following: understanding the reliability of psychiatric diagnosis (see **Topic 9B** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch09lev1sec7#ch09box2), Behavioral and Observational Assessment), choosing appropriate tests for competency (see the topic below, Forensic Applications of Assessment), and comprehending risk–benefit analysis (see **Topic 4A** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch04#ch04box1), Basic Concepts of Validity).

Legislative Sources of Law

In addition to constitutional sources, laws also emanate from the actions of state and federal legislative bodies. These laws are called statutes and are codified by subject areas into codes. For example, the laws passed by Congress at the federal level are codified into 50 topics identified as Title 1 through Title 50 with each area devoted to a specific theme. Three examples include Title 18, Crimes and Criminal Procedure; Title 20, Education; and Title 29, Labor. Each titled area is further subdivided. For example, Title 20, Education, is gargantuan. It consists of 77 chapters, a few of them *hundreds* of pages in length. This includes Chapter 70, Strengthening and Improvement of Elementary and Secondary Schools, in which literally hundreds of specific statutes passed over the last few decades have been collated and cross-referenced. For example, one federal statute mandates that school systems must show adequate yearly progress in order to be eligible for further federal funding. The law further stipulates that "adequate yearly progress" shall be defined by the State in a manner that

> **(i)** applies the same high standards of academic achievement to all public elementary school and secondary school students in the State;
> **(ii)** is statistically valid and reliable;
> **(iii)** results in continuous and substantial academic improvement for all students;
> **(iv)** measures the progress of public elementary schools, secondary schools and local educational agencies and the State based primarily on the academic assessments described in paragraph (3);
> **(v)** includes separate measurable annual objectives for continuous and substantial improvement for each of the following:
>
> (I) The achievement of all public elementary school and secondary school students.
> (II) The achievement of
>
> (aa) economically disadvantaged students;
> (bb) students from major racial and ethnic groups;
> (cc) students with disabilities; and
> (dd) students with limited English proficiency;
>
> except that disaggregation of data under sub-clause (II) shall not be required in a case in which the number of students in a category is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student. (U.S. Code, Title 20, Chapter 70, **http://uscode.house.gov** (http://uscode.house.gov))

As can be seen, legal codes are written with such specificity that their intention cannot easily be overlooked or bypassed. The preceding sample is just one small snippet of law—barely discernible in a vast ocean of literally hundreds of pages of edicts that impact educational practices. But it is clear that these legislative rulings influence psychological testing. For example, in the preceding excerpt, an inescapable inference is that school systems must

use standardized educational achievement tests with established reliability and validity—or else they risk losing federal funds.

Legislatures cannot possibly oversee the implementation of all the statutes they enact. Consequently, it is increasingly common for these bodies to delegate rule-making authority to agencies within the executive branch of government. For example, the U.S. Congress has passed several laws designed to prohibit discrimination in employment. But the enforcement of these laws is left to the Equal Employment Opportunity Commission (EEOC). The following federal laws bear, at least in part, on job discrimination:

- *Civil Rights Act of 1964*, which prohibits employment discrimination based on race, color, religion, gender, or national origin
- *Equal Pay Act of 1963*, which protects women (and men) who perform equal work in the same organization from gender-based wage discrimination
- *Age Discrimination in Employment Act of 1967*, which protects individuals who are 40 years of age or older
- *Americans with Disabilities Act of 1990*, which prohibits employment discrimination against qualified individuals with disabilities in both government and the private sector
- *Rehabilitation Act of 1973*, which prohibits discrimination against qualified individuals with disabilities who work in the federal government
- *Civil Rights Act of 1991*, which authorizes monetary damages in cases of intentional employment discrimination

The EEOC is the federal agency in charge of the administrative and judicial enforcement of the civil rights laws listed earlier. We discuss this important regulatory body in further detail later.

Judicial Sources of Law

Another source of law is the judiciary, specifically, the federal courts and the United States Supreme Court. Indirectly, these bodies make law in several ways. First, they have the authority to review all federal legislative edicts to determine their constitutionality and interpretation. In addition, they can appraise the constitutional validity of any state law, whether constitutional, statutory, or regulatory in origin. In doing so, they have the opportunity to sharpen the focus of laws promulgated by these other sources. For example, in ruling on the constitutionality of state civil commitment laws, federal courts not only have found them unconstitutional, but they have also used this opportunity to publish permissible criteria and procedures for commitment (as discussed previously in relation to the Fourteenth Amendment). The courts also hear lawsuits filed on behalf of individuals or groups. In these cases, court rulings can establish new law. Finally, the courts can make law when the original sources such as constitutional laws or legislative statutes are silent on an important issue:

> In performing their interpretive function, courts will first look at the plain words of any relevant constitutional provision, statute, or regulation and then review the legislative history of a given law, including statements made by the law's sponsors or during committee or public hearing sessions. But if neither of these sources is helpful, or if no relevant law exists, the courts themselves must devise principles to govern the case before them. The principles articulated by courts when they create law are collectively known as common law, or judge-made law. **(Melton et al., 1998 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)** , p. 29)

Typically, common law is conservative, based to the extent possible on the precedent of past cases, rather than created at the whim of the judiciary.

In sum, there are several sources of law: state and federal constitutions, legislative statutes, regulations enacted by agencies such as the EEOC, and judicial interpretations from federal courts and the Supreme Court. These are the primary sources of law that might intersect with the practice of psychological testing. Other sources of law include presidential executive orders and international law, which we do not discuss here because they rarely impact psychological practice.

Now that the reader has an understanding of how, why, and where laws originate, we turn to a review of particular laws that impact the practice of psychological assessment. We partition the discussion into three topics: legal influences on psychological testing in school systems, disability assessment and the law, and legal issues in employment testing. The division is somewhat artificial; for example, the assessment of learning disability—greatly impacted by law—involves both the practice of testing in school systems and the assessment of disability.

12.2 TESTING IN SCHOOL SYSTEMS AND THE LAW

The law has impacted school-based testing in two broad ways: (1) Federal legislation has mandated specific practices in the assessment of students, especially those with disabilities; and (2) lawsuits have shaped and reshaped particular testing practices in school systems over the last 60 years. We will discuss legislative influences in the next section on disability assessment and the law. Our goal here is to provide an overview of influential lawsuits that have molded testing practices in the schools. In the main, these lawsuits have assailed the use of tests, especially in special education placement and as a requirement for high school graduation.

Attacks on cognitive testing in school systems have been with us for a long time. Beginning in the 1960s, these attacks took a new form: lawsuits filed by minority plaintiffs seeking to curtail or ban the use of school-based cognitive tests, especially intelligence tests. In this section we will review the major court cases, summarized in **Table 12.1** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec2#ch12tab1). Later, we will discuss the implications of court decisions for the contemporary use of cognitive tests in schools.

Many of the legal assaults on testing have arisen from the controversial practice of using cognitive test results for purposes of assigning low-functioning students to "vocational" school tracks or to special classes for educable mentally retarded (EMR) persons. Invariably, minority children are assigned to these special tracks and classes in surprising disproportion to their representation in the school population. For example, a typical finding is that minority children are two to three times more likely to be classified as EMR than white children (**Agbenyega & Jiggetts, 1999** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib10)). In a school system comprised of 25 percent minority students, this could translate to EMR classes with about 50 percent minority student representation.

Therein lies the crux of the legal grievance, for special education classes are equated by many with inferior education. Written two decades ago, these observations still hold true:

TABLE 12.1 Major Legal Landmarks in School-Based Cognitive Testing

1967 *Hobson v. Hansen*

Court ruled against the use of group ability tests to "track" students on the grounds that such tests discriminated against minority children.

1970 *Diana v. State Board*

Court ruled against traditional testing procedures for educable mentally retarded (EMR) placement of Mexican American children; State Board of Education enacted special provisions for testing minority children (e.g., bilingual assessment).

1979 *Debra P. v. Turlington*

Court did not rule against the use of a minimum competency test as a condition for high school graduation—a test with excessive failure rate for African American students—but did suspend its use for four years, as a means of providing due process about notification of the new requirement.

1979 *Larry P. v. Riles*

Court ruled that standardized IQ tests are culturally biased against African American children for EMR evaluation and stipulated that the proportion of African American children in these classes must match their proportion in the school population.

1980 *PASE v. Hannon*

In complete contradiction to the *Larry P. v. Riles* decision, the court ruled that standardized IQ tests are not racially or culturally biased.

1984 *Georgia NAACP v. Georgia*
Court ruled that traditional procedures of evaluation do not discriminate against African American children; court also rejected the view that disproportionate representation in EMR classes constituted proof of discrimination.

1994 *Crawford v. Honig*
The judge in the *Larry P. v. Riles* case overruled his earlier ruling so as to allow the use of a standardized IQ test for the evaluation of African American students diagnosed with learning disability.

2000 *GI Forum v. Texas Education Agency*
Court ruled that the use of the Texas Assessment of Academic Skills as part of a high school graduation requirement was permissible despite high failure rates of African American and Latino students.

> If special education actually worked, which it does not, and minority children assigned to EMR classes in the primary grades eventually reached the same level of reading and math achievements as children in regular classrooms, I doubt whether the plaintiffs in these cases would have brought suit. A major problem in the educational system is that special education, even with smaller classes and better trained teachers, still does not work to bring such children up to par. Rather, special education classes perpetuate educational disadvantage. **(Scarr, 1987 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1443))**

Something is amiss in education when well-intentioned placement policies inadvertently perpetuate a legacy of mistreatment of minorities. The legal challenges to school-based testing are certainly understandable, even though sometimes misplaced. After all, the problem is not so much with the tests—which assess academically relevant skills with reasonable validity—but with educational policies that isolate low-functioning students to inefficient placements. Even experts sympathetic to the lawsuits acknowledge that tests often are quite useful, so it is worth examining why killing the messenger has been a popular response to concerns about discriminatory placements.

Hobson v. Hansen (1967)

The first major court case to challenge the validity of ability tests was *Hobson v. Hansen* (1967). In that landmark case, plaintiffs argued that the allocation of financial and educational resources in the Washington, DC, public school system favored white children and, therefore, discriminated against minority children. Among the issues addressed in the trial was the use of standardized group ability tests such as the Metropolitan Readiness and Achievement Test and the Otis Quick-Scoring Mental Ability Test to "track" students according to ability. Children were placed in honors, regular, or basic tracks according to ability level on the tests. One consequence of this tracking method was that minority children were disproportionately represented in the lowest track, which focused on skills and preparation for blue-collar jobs. Placement in this track virtually ruled out entrance to college and entry to a well-paying profession.

Judge Skelly Wright decided the *Hobson* case in 1967, ruling against the use of a tracking system based on group ability tests. Most commentators view his banishment of ability testing for tracking purposes as justified. However, there is good reason to worry about the further implications of Judge Wright's decision, which implied that acceptable tests must measure children's innate capacity to learn. Bersoff (1984) commented on the *Hobson* decision as follows:

> *Hobson*, when read in its entirety, represents the justified condemnation of rigid, poorly conceived classification practices that negatively affected the educational opportunities of minority children and led to permanent stigmatization of blacks as unteachable. But swept within *Hobson's* condemnation of harmful classification practices were ability tests used as the sole or primary decision-making devices to justify placement. Not only was ability grouping as then practiced in the District of Columbia abolished, but tests were banned unless they could be shown to measure children's innate capacity to learn.

Not even ardent hereditarians believe that tests solely measure innate ability. No test could ever pass the criterion mandated by this case.

The *Hobson* case concerned group ability tests and had no direct bearing on the use of individual intelligence tests in school systems. However, it did portend an increasing skepticism about the use of any test—whether group or individual—for purposes of educational placement.

Diana v. State Board of Education (1970)

In *Diana v. State Board of Education* (1970), plaintiffs questioned the use of individual intelligence tests (the WISC and Stanford-Binet) for purposes of placing Mexican American schoolchildren in classes for educable mentally retarded (EMR) persons. *Diana* was a class action suit filed on behalf of nine Mexican American elementary school children who had been placed in EMR classes. The placements were based on individual IQ tests administered by a non-Spanish-speaking psychometrist. When retested in English and Spanish, eight of these nine children showed substantial—sometimes huge—increases in IQ and were, therefore, removed from EMR classes. Faced with this evidence, the California State Board of Education decided to enact a series of special provisions for the testing of Mexican American and Chinese American children. These provisions included the testing of minority children in their primary language, elimination of certain vocabulary and information items that minority children could not be expected to know, retesting of minority children previously placed in EMR classes, and development of new tests normed on Mexican American children. These provisions answered the concerns of plaintiffs, eliminating the need for further court action.

Debra P. v. Turlington (1979)

This was a class action lawsuit filed on behalf of all African American students in Florida against Ralph Turlington, the state Commissioner of Education. At issue was the use of the State Student Assessment Test-Part 2 (SSAT-II), a functional literacy test, as one requirement for awarding a high school diploma. In the 1970s, Florida was one of the states at the forefront of the functional literacy movement. Functional literacy has to do with practical knowledge and skills used in everyday life. A **test of functional literacy (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss330)** might require students to:

- Calculate the balance of a personal checking account when given the starting balance, deposits, withdrawals, and service charges
- Follow simple written directions and instructions in printed materials
- Complete an application form for employment, driver's license, or training program
- Spell basic and useful words correctly (e.g., address, employer, postage, salary, vehicle)
- Comprehend essential abbreviations (e.g., apt., CPU, hwy., M.D., Mr., Rx, SSN)
- Know the meanings of vital words (e.g., antidote, bus stop, caution, exit only, one way, zip code)
- Write a paragraph that is reasonably grammatical and coherent

Currently, about 20 states use a functional literacy test of this genre as one condition of awarding the high school diploma.

However, in Florida in the late 1970s, African American students failed the functional literacy test at a substantially higher rate than white students. Plaintiffs argued the SSAT-II was unfair because African American students received inferior education in substantially segregated schools. The purpose of the lawsuit was to void the use of the test as a requirement for graduation. The information in the following discussion was retrieved from the appeals court decision (*Debra P. v. Turlington*, U.S. Court of Appeals for the Eleventh Circuit, April 27, 1984).

With practical finesse, the court decision offered something to both sides, although state officials likely were happier with the outcome than were the plaintiffs. The nature of the ruling also revealed admirable sensitivity to issues of test validity and psychological measurement on the part of the court. Based on the reasonable belief that a high school diploma should signify functional literacy, the state was permitted to use the test as a diploma requirement. However, the court delayed implementation of the new diploma testing program for four years. This delay served two purposes. First, it provided due process to current students (and their parents), alerting them that a new requirement was being set in place. Second, it gave the state time to prove that the SSAT-II was a fair test of that which is taught in Florida's classrooms. The court wanted proof of what it called "**instructional validity**

(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss158) ." Put simply, the court wanted assurance that the state was teaching what it was testing.

The state undertook a massive evaluation project to prove instructional validity. The Florida Department of Education hired a consulting firm to conduct a four-part study that included (1) teacher surveys asking expressly if the skills tested by the SSAT-II were taught; (2) administrator surveys to demonstrate that school districts utilized remedial programs when appropriate; (3) site visits to verify all aspects of the study; and (4) student surveys to discern if students perceived they were being taught the skills required on the functional literacy test.

Weighing all the evidence carefully over a period of several years, the court ruled that the State of Florida could deny diplomas to students who had not yet passed the SSAT-II, beginning with the class of 1983. Furthermore, the court concluded that the use of the SSAT-II actually helped to mitigate the impact of vestiges of school segregation by motivating students, teachers, and administrators toward a common goal:

> The remarkable improvement in the SSAT-II pass rate among black students over the last six years demonstrates that use of the SSAT-II as a diploma sanction will be effective in overcoming the effects of past segregation. Appellants argue that the improvement has nothing to do with diploma sanctions because the test has not yet been used to deny diplomas. However, we think it likely that the threat of diploma sanction that existed throughout the course of this litigation contributed to the improved pass rate, and that actual use of the test as a diploma sanction will be equally, if not more, effective helping black students overcome discriminatory vestiges and pass the SSAT-II. Thus, we affirm the finding that use of the SSAT-II as a diploma sanction will help remedy vestiges of past discrimination. (U.S. Court of Appeals for the Eleventh Circuit, April 27, 1984)

In sum, the case of *Debra P. v. Turlington* appears to confirm that functional literacy testing can play a constructive role in secondary education.

Larry P. v. Riles (1979)

The case of *Larry P. v. Riles* raised concerns about the use of intelligence tests for assigning African American children to EMR special education classes. In November 1971 attorneys representing several San Francisco families filed for a preliminary injunction seeking to prohibit the use of traditional IQ tests for EMR placement of African American children. The specific grievance was that six African American children in the San Francisco school district had been inappropriately placed into "dead-end" EMR classes based on scores from IQ tests said to be racially and culturally biased against African Americans. As a consequence of this placement it was alleged that the children had suffered irreparable harm. The plaintiffs sought a ban on the use of "culturally biased" IQ tests, asked for reevaluation of all African American EMR children, requested special assistance for those who returned to the regular classroom, and sought a quota limiting assignment of African American children to EMR classes. The quota was defined in proportion to overall African American representation in the school district population.

In 1972 Judge Robert Peckham granted a preliminary injunction, restraining school officials in San Francisco from placing primary reliance on IQ tests in EMR placements for African American children. He also ordered that African American EMR children should be reevaluated and that those who were returned to regular classes should be given special help. However, he was wary of the plaintiffs' proposed ratio system limiting African American enrollment in EMR classes.

The case of *Larry P.* eventually went to trial in 1978. More than 50 expert witnesses were called and over 200 reports, studies, and exhibits were received in evidence. In the end the plaintiffs prevailed. In 1979 Judge Peckham ruled that individual intelligence tests "are racially and culturally biased, have a discriminatory impact against black children, and have not been validated for the purpose of essentially permanent placements of black children into educationally dead-end, isolated, and stigmatizing classes for the so-called educable mentally retarded."

This decision was based, in part, on certain assumptions about the nature of intelligence that are not necessarily shared by experts in the field. For example, after reviewing the trial transcript—some ten thousand pages in length —Elliott (1987) concluded that the legal opinion in *Larry P.* was based on the following assumptions: that

intelligence is the innate ability to learn, that a culturally fair test should measure innate ability, and that a culturally fair test should produce equal scores for all relevant subgroups. If these assumptions are correct, then the legal opinion cited in *Larry P.* follows with inexorable logic. However, very few assessment specialists embrace the antiquated view that it is meaningful or useful to define intelligence as innate ability to learn.

Within California, the decision effectively abolished the use of individual intelligence tests for placement of African American students in EMR classes. In 1984 the decision was affirmed by the U.S. Ninth Circuit Court of Appeals, and in 1986 the ban was extended so that IQ tests could not be used for *any* special education placement of African American children in the public schools of California.

Although it is arguable whether the *Larry P.* decision was good social science, there is no denying the profound policy implications of this case:

> For special education, the negative results are reduced precision and objectivity of assessment, reduced precision of placement, reduced morale of and faith in the professionals charged with assessment, some downgrading of the once-central importance of developing intellectual skills, and reduced services for slow-learning, non-LD children in the 65–80 range. The positive results are broader and newer kinds of assessment (if there is time for the breadth, and norms for the novel tests) and some fresh thinking about programs for children having difficulty in school. (Elliott, 1987)

One major consequence of *Larry P.* has been a huge reduction in the number of children assigned to self-contained EMR classes. For example, in California the number of EMR children went from a high of 58,000 in 1968–1969 to approximately 13,000 in 1984. For some mildly retarded children, alternative placement in regular classrooms has been beneficial, but for others who are now not eligible for any special help, the aftermath of court-influenced placement policies is more questionable (**Powers & Hagans-Murillo, 2004 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1319)**).

Parents in Action on Special Education (PASE) v. Joseph P. Hannon (1980)

PASE v. Hannon was litigated in 1980, just one year after the landmark *Larry P. v. Riles* case. In this suit, attorneys for two African American student plaintiffs argued that the children were inappropriately placed in educable mentally handicapped (EMH) classes because of racial bias in the IQ tests used for placement. The case was tried as a class action suit, meaning that the plaintiffs represented the category of all similar children in Chicago. Even though the issues in the *PASE* class action suit were substantially the same as the preceding case, the presiding judge came to exactly the opposite conclusion. Judge John Grady ruled that intelligence tests are not culturally biased against African American children.

Astonishingly, in his written opinion Judge Grady commented on the cultural fairness of every single item on the WISC, WISC-R, and Stanford-Binet, finding all but 9 of the 488 items to be culturally fair. He concluded that the 9 biased items were not sufficient in number to render the tests discriminatory, and he endorsed their ongoing use for evaluation of minority children. Although little has been made of the judge's transgression, it would be considered a colossal breach of professional ethics were a psychologist to publish individual test items in the public record.

Georgia NAACP v. Georgia (1984)

In this case the NAACP alleged that evaluation procedures used in the state of Georgia discriminated against African American children, resulting in their overrepresentation in EMR classes. However, the U.S. Court of Appeals ruled in 1984 that discrimination did not exist. Furthermore, the court rejected the notion that overrepresentation of African American children in EMR classes was a sufficient basis to prove discrimination.

Crawford v. Honig (1994)

This case initiated a reexamination of the rights of minority children in special education in California. Contrary to other cases in which the lawyers and parents of minority children asked for a ban on the use of traditional tests, the purpose of *Crawford v. Honig* was exactly the opposite—to obtain legal permission for using tests such as the Wechsler Intelligence Scale for Children-Revised (WISC-R) with African American children. The case was filed by

the parents of Demond Crawford, an African American student diagnosed with learning disability. His parents understood the value of standardized intelligence tests in the assessment of learning disability and wanted school psychologists to use these traditional instruments in their evaluation. However, as a direct consequence of the *Larry P. v. Riles* decision, it was *illegal* in 1994 for psychologists to administer the WISC-R, or any other mainstream IQ test, to African American children in California, even with the permission of the parents. A psychologist who did so risked fines and jail time for breaking the law. In this lawsuit, Judge Robert Peckham, the same judge who presided in the *Larry P. v. Riles* case, overruled his earlier finding so as to permit the use of standardized IQ tests in the evaluation of African American children upon the formal request of their parents. This is an excellent example of the fact that laws can be reshaped in response to changing social conditions.

GI Forum v. Texas Education Agency (2000)

In this court suit, filed on behalf of seven African American and Latino high school students in Texas, plaintiffs challenged the use of the Texas Assessment of Academic Skills (TAAS) as a requirement for high school graduation on the grounds that it discriminated unfairly against minority students and violated their right to due process. They pointed out that substantial disparities in resources existed between "white" schools—those with a preponderance of white students—and minority schools—those with a preponderance of minority students. In the view of plaintiffs, this was the explanation for the differential failure rates. In fact, 67 percent of African American, 59 percent of Latino, and 31 percent of white students failed the exam the first time it was used in 1991.

After hearing expert witnesses over many months, the court ruled in favor of state education officials, citing several compelling reasons. Although the court agreed with plaintiffs that disparities in resources did exist, it found no evidence that these inequalities caused the higher failure rate of minority students. The court also pointed out that the TAAS was constructed with great care and possessed "curricular validity"; that is, it tested what was actually taught. This quality of a test is the same thing as instructional validity, as described earlier in *Debra P. v. Turlington*. Officials also noted that the TAAS was just one condition of awarding the diploma, not the sole factor; attendance, passing grades, and completion of the required curriculum also are needed. The court praised the humane manner of test implementation, noting that students first encounter the TAAS in the tenth grade and are provided remedial courses for any of the three subsections (reading, math, writing) that they fail. The cutoff score of 70 percent for each curricular area was deemed reasonable. Moreover, the court noted, students have a minimum of *seven* additional opportunities to pass the test. Finally, the court found it "highly significant that minority students have continued to narrow the passing rate gap at a rapid rate." Similar to the findings in *Debra P. v. Turlington*, this case demonstrated that a well-designed graduation test can be an engine of positive social change.

12.3 DISABILITY ASSESSMENT AND THE LAW

Individuals with disabilities are afforded many legal protections, some of which impact the use of psychological tests. In this section, we review two broad areas in which legislation has been written to defend individuals with disabilities: school-based assessment of children with disabilities, and employment-based testing of persons with disabilities. The coverage is purposefully brief. Readers can find lengthier discussions in Bruyere and O'Keeffe (**1994** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib228)**), Salvia and Ysseldyke (**2001** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1432)**), and Stefan (**2001** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1550)**).

Public Law 94-142

In 1975, the U.S. Congress passed a compulsory special education law, **Public Law 94-142** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss259)** , known as the Education for All Handicapped Children Act.[1] **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec3#ch12fn01)** According to Ballard and Zettel (**1977** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib80)**), this law was designed to meet four major goals:

- To ensure that special education services are available to children who need them
- To guarantee that decisions about services to disabled students are fair and appropriate
- To establish specific management and auditing requirements for special education
- To provide federal funds to help the states educate disabled students

Many practices in the assessment of disabled persons stem directly from the provisions of Public Law 94-142. For example, the law specifies that each disabled student must receive an individualized education plan (IEP) based on a comprehensive assessment by a multidisciplinary team. The IEP must outline long-term and short-term objectives and specify plans for achieving them. In addition, the IEP must indicate how progress toward these objectives will be evaluated. The parents are intimately involved in this process and must approve the particulars of the IEP.

Pertinent to testing practices, PL 94-142 includes a number of provisions designed to ensure that assessment procedures and activities are fair, equitable, and nondiscriminatory. Salvia and Ysseldyke (1988) summarize key provisions which include assessment in the native language with validated tests administered by trained personnel; appraisal in areas related to the specific disability, including—when appropriate—hearing, vision, emotional functioning, academic performance, communication skills, motor skills, and general intelligence; and, evaluation by a multidisciplinary team that includes a teacher or specialist with knowledge of the area of suspected disability.

PL 94-142 also contains a provision that disabled students should be placed in the least restrictive environment—one that allows the maximum possible opportunity to interact with nonimpaired students. Separate schooling is to occur only when the nature or the severity of the disability is such that instructional goals cannot be achieved in the regular classroom. Finally, the law contains a due process clause that guarantees an impartial hearing to resolve conflicts between the parents of disabled children and the school system.

In general, the provisions of PL 94-142 have provided strong impetus to the development of specialized tests that are designed, normed, and validated for children with specific disabilities. For example, in the assessment of a child with visual impairment, the provisions of PL 94-142 virtually dictate that the examiner must use a well-normed test devised just for this population rather than relying upon traditional instruments.

Public Law 99-457

In 1986, Congress passed several amendments to the Education for All Handicapped Children Act, expanding the provisions of PL 94-142 to include disabled preschool children. **Public Law 99-457** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss260)** requires states to provide free appropriate public education to disabled children ages 3 through 5. The law also mandates financial grants to

states that offer interdisciplinary educational services to disabled infants, toddlers, and their families, thus establishing a huge incentive for states to serve children with disabilities from birth through age 2. Public Law 99-457 also provides a major impetus to the development and validation of infant tests and developmental schedules. After all, the early and accurate identification of at-risk children would appear to be the crucial first step in effective interdisciplinary intervention.

No Child Left Behind Act

In the context of school-based testing and the law, an important development is the 2001 **No Child Left Behind Act** (NCLB). The ambition of this act was to improve education through standards-based reforms that require states to implement assessments in basic educational skills. NCLB is a complex and far reaching law that expands the federal role in public education. There are important implications for educational and psychological testing in this act. The six elements of the law include:

Annual Testing: The heart of NCLB is yearly testing toward prescribed academic goals, especially in reading and math. States are required to test students in grades 3-8 annually in reading and math, in order to receive federal funds. States also are required to test students in science at least once in elementary, middle, and high school. While each state is free to set its own academic standards, the testing programs must be closely aligned with the state standards.

Academic Progress: Schools are required to make Adequate Yearly Progress (AYP), as defined by formulas provided in NCLB, or they must enact prescribed changes, also defined in the law. The prescribed changes increase in scope over time. For example, if a school fails to make AYP two years in a row, it must create a plan to improve teaching in the deficient topic(s). After four years of not meeting AYP, prescribed options include introducing a new curriculum, extending the school day, and replacement of teachers. After six years, the school faces draconian measures that include being turned over to a private company, management by the state office of education, or closure.

Report Cards: States are required to furnish annual report cards that provide information on student achievement, broken down by subgroups (e.g., minorities, English as a second language students) and also by school districts. Districts must provide similar report cards.

Teacher Qualifications: Every teacher in core content areas must be "highly qualified" in each subject taught. Under the law, this refers to special certifications and demonstrations of proficiency in the subject matter.

Reading First: This component of NCLB created a competitive grant program to aid schools districts in setting up empirically based reading programs for children in grades K-3, with priority given to low-income areas. Congress later cut funding drastically for this element of the act.

Funding Changes: Funding formulas were revised so that school districts with high concentrations of low income children would receive better support and would have greater flexibility in using federal funds.

In the years since its inception, NCLB has remained controversial, and efforts to modify it often make headlines. Whether the act is accomplishing its stated intentions is still an open question. But it is an issue that can be investigated in an empirical, nonpartisan manner. For example, Wang, Beckett, and Brown (**2006 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1715)**) provide an even-handed synthesis of research-based findings on the impact of NCLB, summarizing pros and cons. The general tone of their review is mildly supportive. While citing several problems with NCLB (e.g., failure to provide adequate funding for test development and personnel training, failure to acknowledge genetic and socioeconomic influences), the authors conclude that the law is bringing about positive changes in student learning.

But not all reviewers agree with this optimistic inference. The potential distorting effects of the high-stakes testing dictated by NCLB remain a serious concern. Nichols, Glass, and Berliner (**2006 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1233)**) analyzed longitudinal data from 25 states on the relationship between high-stakes pressure and improvements in student achievement as

measured by the National Assessment of Educational Progress (NAEP). NAEP consists of periodic assessments in mathematics, reading, science, writing, the arts, civics, economics, geography, and U.S. history. The tests are administered uniformly across the nation with the same test booklets. Based on sophisticated correlational analyses across time, Nichols et al. (**2006** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1233)**) found no value in high-states testing. To the contrary, they found the impact to be insidiously negative. Specifically, their analyses revealed that

- States with greater proportions of minority students implement accountability systems that exert greater pressure. This suggests that any problems associated with high-stakes testing will disproportionately affect America's minority students.
- High-stakes testing pressure is negatively associated with the likelihood that eighth and tenth graders will move into 12th grade. Study results suggest that increases in testing pressure are related to larger numbers of students being held back or dropping out of school.
- Increased testing pressure produced no gains in NAEP reading scores at the fourth- or eighth-grade levels (**Nichols et al., 2006** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1233)** , p. 5).

These authors call for a moratorium on policies that force school systems to use high-stakes testing. By implication, this would mean that key elements of NCLB ought to be suspended.

Americans With Disabilities Act

The 1990 **Americans with Disabilities Act** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss10)** (ADA) forbids discrimination against qualified individuals with disabilities in both the public sector (e.g., government agencies and entities receiving federal grants) and the private sector (e.g., corporations and other for-profit employers). Under the ADA, *disability* is defined as a physical or mental impairment that substantially limits one or more of the major life activities (Parry, 1997). Examples of ADA-recognized disabilities include sensory and physical impairments (e.g., blindness, paralysis), many mental illnesses (e.g., major depression, schizophrenia), learning disabilities, and attention-deficit/hyperactivity disorder.

Under the ADA, the process of qualifying an individual for work or educational accommodations requires current, detailed, and professional documentation. For example, a graduate student who was seeking a special arrangement for taking tests (such as a quiet room) because of attentional problems might need to submit a comprehensive endorsement from a licensed psychologist, detailing the history, current functioning, clinical diagnosis of attention-deficit/hyperactivity disorder, and necessity for accommodations (**Gordon & Keiser, 1998** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib621)**). In other words, the ADA is a civil rights act, not a program of entitlement:

> The ADA does not guarantee equal outcomes, establish quotas, or require preferences favoring individuals with disabilities. Rather, the ADA is intended to ensure access to equal employment opportunities based on merit. The ADA is designed to "level the playing field" by removing the barriers that prevent qualified individuals with disabilities from having access to the same employment opportunities that are available to individuals without disabilities. (**Klimoski & Palmer, 1994** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib899)** , p. 45)

In sum, the purpose is to ensure that individuals who are otherwise qualified for jobs or educational programs are not denied access or put at improper disadvantage simply because of a disability.

In regard to psychological testing, an important provision of the ADA is that agencies and institutions must make reasonable testing accommodations for persons with disabilities. With appropriate documentation (discussed earlier), the relevant accommodations might include any of the following:

- Assistance in completing answer sheets

- Audiotape or oral presentation of written tests
- Special seating for tests
- Large-print examinations
- Retaking exams
- Dictating rather than writing test answers
- Printed version of verbal instructions
- Extended time limit

In general, changes in the testing medium (e.g., from written to oral) are consistent with the intention of ADA, if such a change is needed to accommodate a disability. For example, an appropriate accommodation in the testing medium would be the audiotaped presentation of test items for persons who are visually impaired. On the other hand, changing a test from a printed version into a sign language version for persons with hearing impairment would be considered translation into another language, not a simple change of medium.

In most testing accommodations mandated by the ADA, it is necessary to change the time limits, usually by providing extra time. This raises problems of test interpretation, especially when a strict time limit is essential to the validity of a test. For example, Willingham, Ragosta, Bennett, and others (**1988 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1762)**) found that extended time limits on the SAT significantly reduced the validity of the test as a predictor of first-year college grades. This was especially true for examinees with learning disabilities, whose first-year grades were subsequently overpredicted by their SAT scores. Thus, although it seems fair to provide extra time on a test when the testing medium has been changed (e.g., audiotaped questions replacing the printed versions), from a psychometric standpoint, the challenge is to determine *how much* extra time should be provided so that the modified test is comparable to the original version. Nester (**1994 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1225)**) and Phillips (**1994 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1290)**) provide thoughtful perspectives on the range of reasonable accommodations required by the ADA.

Cognitive Disability and the Death Penalty

One way that laws evolve in American society is through decisions of the Supreme Court. In a 2002 court case (*Atkins v. Virginia*), the Supreme Court held that the execution of mentally retarded convicts is "cruel and unusual punishment" prohibited by the Eighth Amendment. In speaking for the 6-3 majority, Chief Justice John Paul Stevens wrote:

> We are not persuaded that the execution of mentally retarded criminals will measurably advance the deterrent or the retributive purpose of the death penalty. Construing and applying the Eighth Amendment in the light of our "evolving standards of decency," we therefore conclude that such punishment is excessive and that the Constitution "places a substantive restriction on the State's power to take the life" of a mentally retarded offender. (*Atkins v. Virginia*, 2002, p. 321)

This new constitutional standard has profound implications, literally of life and death, for the proper application of psychological tests with persons who display intellectual disability. Choosing the appropriate tests, getting the results right, and offering an accurate diagnosis of intellectual disability could determine whether some examinees face death row.

This was certainly relevant for Doil Lane, who was convicted of the heinous rape and murder of a nine-year-old girl and sentenced to death, principally on his confession (DNA testing was inconclusive). This confession of a highly suggestible young man with intellectual disability may have been false. Whether or not his confession was true, there is no question as to presence of significant intellectual disability:

> As a child, he spent years as a resident of a special school in Texas for mentally disabled students. His I.Q. has tested between 62 and 70. His mental deficiencies are so obvious that the report by the Kansas police officer who first interviewed him noted Lane seemed "mentally retarded." The former chief psychologist of the Texas Division of Criminal Justice assessed his intelligence in 1998 and concluded he had mental retardation. When his police interrogation was over, Lane—a thirty-year-old—climbed into the interrogating officer's lap. At his

trial in Texas, Lane asked the judge for crayons so that he could color pictures. The judge denied the request. **(Human Rights Watch, 2001**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib795) , p. 38)

In response to the Supreme Court decision, Texas Governor Rick Perry commuted the death sentence of Doil Lane to life in prison.

[1]Each congressional law receives two numbers, one referring to the particular Congress that passed it, the other referring to the law itself. Thus, Public Law 94-142 is the 142nd law passed by the 94th Congress.

12.4 LEGAL ISSUES IN EMPLOYMENT TESTING

Nearly every aspect of the employment relationship is subject to the law: recruitment, screening, selection, placement, compensation, promotion, and performance appraisal all fall within the domain of legal interpretations **(Cascio, 1987 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib279))**. However, courts and legislative bodies have reserved special scrutiny for employment-related testing. The practitioner who refuses to learn relevant legal guidelines in personnel testing does so at great peril, because unwise practices can lead to costly and time-consuming litigation **(Case Exhibit 12.1 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec4#ch12box2))**.

CASE EXHIBIT 12.1

Unwise Testing Practices in Employee Screening

According to the Associated Press of July 11, 1993, the Target discount chain agreed to settle out of court in a class-action lawsuit filed on behalf of an estimated 2,500 job applicants. Prospective security guards for Target were required to take the Rodgers Psychscreen, a 704-item condensed combination of the CPI and the MMPI. Several applicants objected to answering the test, which included questions about God, sex, and bowel movements. Target agreed to pay $1.3 million, including $60,000 to four plaintiffs named in the lawsuit. Although Target admitted no wrongdoing in the case, corporate officers agreed not to use the Psychscreen test for at least five years.

Sibi Soraka was one of the plaintiffs in the lawsuit. He found the questions to be "off-the-wall and bizarre." He claimed that the cumulative effect of answering the questions made him palpably ill. He added: "It doesn't take Einstein to figure out that these questions really don't have any bearing on our world and life today, or certainly on a job walking around looking for shoplifters." Target corporation defended the testing practice, noting that Psychscreen is commonly used in the evaluation of law enforcement officers. Attorneys for Soraka disagreed, citing a lack of evidence that the test helped identify good versus poor risks for employment. They noted that about 800 of the 2,500 applicants were denied employment based solely upon Psych-screen results.

This case illustrates that the psychometric soundness of an instrument is not the only criterion in test selection. In addition, test users must show that the instrument is relevant to their application. Furthermore, issues of acceptability to prospective examinees must be considered.

Personnel testing is particularly sensitive because the consequences of an adverse decision are often grave: The applicant does not get the job, or an employee does not get the desired promotion or placement. Recognizing that employment testing performs a sensitive function as gatekeeper to economic advantage, Congress has passed laws sharply regulating the use of testing. The courts have also rendered decisions that help define unfair test discrimination. In addition, regulatory bodies have published guidelines that substantially impact testing practices. We will provide a current perspective on the regulation of personnel testing by tracing the development of laws, regulations, and major court cases.

It may surprise the reader to learn that employment testing has raised legal controversy only in the last 35 years **(Arvey & Faley, 1988 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib63))**. During this period, several definitive court decisions and path breaking governmental directives have helped define current legal trends. These landmarks are depicted in **Table 12.2 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec4#ch12tab2)** , beginning with the Civil Rights Act of 1964, proceeding through the federal regulations of the Equal Employment Opportunity Commission (EEOC), and concluding with very recent court cases and legislative developments. We will review these landmarks in chronological order.

TABLE 12.2 Major Legal Landmarks in Employment Testing

1964 *Myart v. Motorola.* This case set the precedent for courts to hear employment testing cases.

1964 Civil Rights Act. This act prohibits job discrimination based on sex, race, color, religion, or national origin.

1966 EEOC Guidelines. The first published guidelines on employment testing practices.

1971 *Griggs v. Duke Power Company.* The Supreme Court rules that employment test results must have a demonstrable link to job performance.

1973 *United States v. Georgia Power Company.* Ruling strengthens the authority of EEOC guidelines for studies of employment testing validity.

1975 *Albemarle v. Moody.* EEOC guidelines strengthened; subjective supervisory ratings ruled a poor basis for validating tests.

1976 *Washington v. Davis.* Court ruled that performance in a training program was a sufficient basis against which to validate a test.

1978 *Uniform Guidelines on Employee Selection.* These guidelines defined adverse impact by the four-fifths rule and incorporated criteria for validity in employee selection studies.

1988 *Watson v. Fort Worth Bank and Trust.* The court ruled that subjective employment devices such as the interview can be validated; employees can claim disparate impact based on interview-based promotion policies.

1990 Americans with Disabilities Act. This act sharply limits the reasons for not hiring a disabled person. One provision is that medical tests may not be administered prior to an offer of employment.

1991 Civil Rights Act. This act outlaws subgroup norming of employee selection tests.

Early Court Cases and Legislation

During the presidency of Lyndon Johnson, Congress passed the Civil Rights Act of 1964. This early civil rights legislation had a profound effect on employee-testing procedures. In addition to broad provisions designed to prevent discrimination in many social contexts, Title VII of this act prohibits employment practices that discriminate on the basis of race, color, religion, sex, or national origin. The act established several important general principles relevant to employment testing **(Cascio, 1987 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib279))**:

- Discriminatory preference for any group, minority or majority, is barred by the act.
- The employer bears the burden of proof that all requirements for employment, including test scores, are related to job performance.
- Professionally developed tests used in personnel testing must be job related.
- In addition to open and deliberate discrimination, the law forbids practices that are fair in form but discriminatory in operation.
- Intent is irrelevant: the plaintiff need not show that discrimination was intentional.
- In spite of these proscriptions, job-related tests and other measuring devices are deemed both legal and useful.

The 1964 legislation also created the Equal Employment Opportunity Commission (EEOC) to develop guidelines defining fair employee-selection procedures. The initial guidelines, published in 1966, were vague. Later revisions of these guidelines, including the *Uniform Guidelines on Employee Selection* (1978), were quite specific and have been used by the courts to help resolve legal disputes regarding employment-testing practices (see the following section).

The 1964 *Myart v. Motorola* case marked the first involvement of the courts in employment testing. The issues raised by this landmark case are still reverberating today. Leon Myart was an African American applicant for a job at one of Motorola's television assembly plants. Even though he had highly relevant job experience, Mr. Myart was refused a position because his score on a brief screening test of intelligence fell below the company cutoff. Claiming racial discrimination, he filed an appeal with the Illinois Fair Employment Practices Commission. The state examiner found in favor of the complainant and directed that the Motorola company should offer Mr. Myart a job. In addition, the examiner ruled that the particular test should not be used in the future and that any new test

should "take into account the environmental factors which contribute to cultural deprivation." In essence, the examiner concluded that Motorola's employment-testing practices were unfair because they acted as a barrier to the employment of culturally deprived and disadvantaged applicants. Even though the case was later overturned for lack of evidence, *Myart v. Motorola* did set the precedent to hear such complaints in the court system (**Arvey & Faley, 1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib63)).

Advent of EEOC Employment Testing Standards

During the 1970s, several court cases helped shape current standards and practices in employment testing. The focus of *Griggs v. Duke Power Company* (1971) was the use of tests—in this case the Wonderlic Personnel Test and the Bennett Mechanical Comprehension Test—as eligibility criteria for employees who wanted to transfer to other departments. In particular, employees at Duke Power Company who lacked a high school education could qualify for transfer if they scored above the national median on both tests. This policy appeared to discriminate against African American employees since it was disproportionately difficult for them to gain eligibility for transfer. However, lower courts found no discriminatory intent and therefore found in favor of the power company.

In 1971, the Supreme Court reversed the lower court findings, ruling against the use of tests without their validation. The decision emphasized several points of current relevance (**Arvey & Faley, 1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib63)):

- Fairness in employment testing is determined by consequences, not motivations.
- Testing practices must have a demonstrable link to job performance.
- The employer has the burden of showing that an employment practice such as testing is job related.
- Diplomas, degrees, or broad testing devices are not adequate as measures of job-related capability.
- The EEOC testing standards deserve considerable deference from employment testers.

These employment testing guidelines were further refined in a 1973 court decision, *United States v. Georgia Power Company.* In this case, the Georgia Power Company presented a validation study to support its employment-testing practices when its policies were shown to have an adverse impact upon the hiring and transferring of African Americans. However, the validation study was weak, in part because it was based upon multiple discriminant analysis, a complex statistical technique rarely used for this purpose. The courts ruled that the validation study was inadequate since it did not adhere to EEOC guidelines for evaluating validity studies. This finding ensconced the EEOC guidelines as virtually the law of the land in employment-testing practices.

Several other court cases in the 1970s and 1980s also served to strengthen the authority of EEOC testing guidelines. These cases were quite complex and involved multiple issues in addition to those cited here. In *Albemarle v. Moody* (1975), the Supreme Court deferred to EEOC guidelines in finding that subjective supervisory ratings are ambiguous and, therefore, constitute a poor basis for evaluating the validity of an employment selection test. The central issue in *Washington v. Davis* (1976) was whether performance in a training program (as opposed to actual on-the-job performance) was a sufficient basis for determining the job-relatedness of the employment selection procedures. In this case, the Supreme Court ruled that performance in a police officer training program was a sufficient criterion against which to validate a selection test.

In *State of Connecticut v. Teal*, the U.S. Supreme Court sided with four African American state employees who had failed a written test that was used to screen applicants for the position of welfare eligibility supervisor. The workers claimed unfair discrimination, noting that only 54 percent of minority applicants passed, compared to 80 percent for whites. In its defense, the state of Connecticut argued that discrimination did not exist, since 23 percent of the successful African American applicants were ultimately promoted, compared to 14 percent for whites. The Court was not impressed with this argument, noting that Title VII of the 1964 Civil Rights Act was specifically designed to protect individuals, not groups. Thus, any unfairness to an individual is unacceptable. Further analysis of fair employment court cases can be found in Arvey and Faley (**1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib63)), Cascio (**1987** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib279)), Kleiman and Faley (**1985** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib897)), and Russell (1984).

Uniform Guidelines on Employee Selection

During the 1970s, several federal agencies and professional groups proposed revisions and extensions of the existing EEOC employment testing guidelines. The revisions were developed in response to court decisions that had interpreted EEOC guidelines in a narrow, inflexible, legalistic manner. However, the existence of several sets of competing guidelines was confusing, and strong pressures were exerted upon the involved parties to forge a compromise. These efforts culminated in a consensus document known as the 1978 *Uniform Guidelines on Employee Selection.*

The *Uniform Guidelines* quickly earned respect in court cases and were frequently cited in the resolution of legal disputes. The new guidelines contain interpretation and guidance not found in earlier versions, particularly regarding adverse impact, fairness, and the validation of selection procedures, as discussed later.

The *Uniform Guidelines* provide a very specific definition of adverse impact. In general, when selection procedures favor applicants from one group (usually males or whites), the basis for selection is said to have an **adverse impact (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss04)** on other groups (usually females or nonwhites) with a lower selection proportion. The *Uniform Guidelines* define adverse impact with a four-fifths rule. Specifically, adverse impact exists if one group has a selection rate less than four-fifths of the rate of the group with the highest selection rate. For example, consider an employer who has 200 applicants in a year, 100 African American and 100 white. If 120 persons were hired, including 80 whites and 40 African Americans, then the percentage of whites hired is 80 percent (80/100), whereas the percentage of African Americans hired is 40 percent (40/100). Since the selection rate for African Americans is only half that of whites (40 percent/80 percent), the employer might be vulnerable to charges of adverse impact. We should note that the *Uniform Guidelines* suggest caution about this rule when sample sizes are small.

The *Uniform Guidelines* also pay more attention to fairness than previous documents. Fairness is treated in the following manner:

> When members of one racial, ethnic, or sex group characteristically obtain lower scores on a selection procedure than members of another group, and the differences are not reflected in differences in a measure of job performance, use of the selection procedure may unfairly deny opportunities to members of the group that obtain the lower scores. Furthermore, in cases where two or more selection procedures are equally valid, the employer is obliged to use the method that produces the least adverse impact.

The *Uniform Guidelines* also establish a strong affirmative action responsibility on the part of employers. If an employer finds a substantial disparity in persons hired from a subgroup compared to their availability in the job market, several corrective steps are recommended. These corrective measures include specialized recruitment programs designed to attract qualified members of the group in question, on-the-job training programs so that affected minorities do not get locked into dead-end jobs, and a revamping of selection procedures to reduce or eliminate exclusionary effects.

Finally, the guidelines provide specific technical standards for evaluating validity studies of employee selection procedures. The courts will almost certainly consult these *Uniform Guidelines* if employees bring suit against the company for alleged unfairness in employee selection practices. Thus, it is a foolish employer who does not pay special attention to these technical criteria. For example, one criterion concerns the use of performance scores obtained during training programs:

> Where performance in training is used as a criterion, success in training should be properly measured and the relevance of the training should be shown either through a comparison of the content of the training program with the critical or important work behavior(s) of the job(s), or through a demonstration of the relationship between measures of performance in training and measures of job performance.

Thus, preemployment evaluation of job candidates in a training program may constitute a valid method of employee selection, but only if a strong link exists between the task demands of training and the requirements of the actual job.

The *Uniform Guidelines* contain many other criteria that we cannot review here. We urge the reader to read this fascinating and influential document which is often cited in court cases on employment discrimination.

Legal Implications of Subjective Employment Devices

In many corporations, promotions are based upon the subjective judgment of senior managers. A common practice is for one or more managers to interview several qualified employees and offer a promotion to the one candidate who appears most promising. The selection of this candidate is typically based on subjective appraisal of such factors as judgment, originality, ambition, loyalty, and tact. Until recently, these subjective employment devices appeared to be outside the scope of fair employment practices codified in the *Uniform Guidelines* and other sources.

However, in a civil rights case, *Watson v. Fort Worth Bank and Trust* (1988), the Supreme Court made it easier for employees to prove charges of race or sex discrimination against employers who use interview and other subjective assessment devices for employee selection or promotion. We outline the factual background of this important case before discussing the legal implications (**Bersoff, 1988 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib154)**).

Clara Watson, an African American employee at Fort Worth Bank and Trust, was rejected for promotion to supervisory positions four times in a row. Each time, a white applicant received the promotion. Watson obtained evidence showing that the bank had never had an African American officer or director, had only one African American supervisor, and paid African American employees lower salaries than equivalent white employees. Furthermore, all supervisors had to receive approval from a white male senior vice president for their promotion decisions. The bank did not dispute that it made hiring and promotion decisions solely on the basis of subjective judgment. When an analysis of promotion patterns confirmed statistically significant racial disparities, Watson brought suit against the bank.

Two legal theories were available for Watson to litigate her claim under Title VII of the 1964 Civil Rights Act. The two theories are called "disparate treatment" and "disparate impact." A disparate treatment case is more difficult to litigate, since the plaintiff must prove that the employer engaged in intentional discrimination. In a disparate impact case, intention is irrelevant. Instead, the plaintiff need merely show that a particular employment practice—such as using a standardized test—results in an unnecessary and disproportionately adverse impact upon a protected minority.

The lower courts ruled that Watson was restricted to the more limited disparate treatment approach since the employer had used subjective evaluation procedures. Furthermore, the lower courts ruled that the bank had not engaged in intentional discrimination and did have legitimate reasons for not promoting Watson. Nonetheless, the Supreme Court agreed to hear the case in order to determine whether a disparate impact analysis could be applied to subjective employment devices such as interview. Relying heavily upon a brief from the American Psychological Association (**APA, 1988 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib28)**), the Supreme Court ruled unanimously that the disparate impact analysis is applicable to subjective or discretionary promotion practices based on interview. In effect, the Court ruled that subjective employment devices such as interview can be validated. Thus, employers do not have unmonitored discretion to evaluate applications for promotion based on subjective interview. As a consequence of *Watson v. Fort Worth Bank and Trust,* employers must be ready to defend all their promotion practices—including subjective interview—against claims of adverse impact.

Recent Developments in Employee Selection

Recent court cases also have impacted personnel testing. The issue in *Soraka v. Dayton Hudson* was whether corporations can use a personality test as a basis for preemployment screening for mental health problems in job applicants. As discussed previously, Soraka was required to take the Rodgers Psychscreen as part of the application process for a position as security guard. The Psychscreen is a true-false personality inventory intended to identify persons with psychological problems such as depression and anxiety. Soraka filed suit against the department store, claiming that individual questions about his sexual practices and religious beliefs were a

violation of his civil rights. This case was interesting because it pertained to the value and validity of individual items as opposed to overall test scores. The courts have long held that preemployment testing must have demonstrated relevance to job performance or it cannot be used. However, the courts have not required validity evidence for individual test items. Soraka won his case, which was appealed by Dayton Hudson. In 1993, the company settled out of court. This litigation is summarized in **Case Exhibit 12.1** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec4#ch12box2)** found earlier in this section.

Another recent court case illustrates how litigation will continue to clarify the scope of ADA in regard to psychological testing. In *Karraker v. Rent-A-Center* (2005), a federal appeals court unanimously invalidated the use of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) as a job screening test, citing ADA restrictions on preemployment medical tests. The defendants argued in vain that their use of the test was solely to measure traits of character and personality such as honesty, preferences, and reliability—all legal under ADA. The appeals court held that the MMPI-2 was designed, at least in part, to reveal mental illness. As such, the effect of using the test was to hurt employment prospects for individuals with a mental disability, a direct violation of ADA.[2] **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec4#ch12fn02)** The defendants paid a substantial sum to settle a class action suit filed by employees and agreed to stop using the test in California.

[2]Oddly enough, in one of those twists so typical of how law is interpreted, it appears that the MMPI-2 still can be used legally in employment settings if the employer makes a *conditional* offer of employment before requiring that candidates take the test.

12.5 FORENSIC APPLICATIONS OF ASSESSMENT

Psychology and the legal system have had a long and uneasy alliance characterized by mistrust on both sides. Within the legal system, lawyers and judges maintain antipathy toward the testimony of psychologists because of a concern that their opinions are based upon "junk science" (or perhaps no science at all) and also because of a belief (not entirely unfounded) that some expert witnesses will profess almost any viewpoint that serves the interests of a defendant. Within the mental health profession, psychologists find the adversarial aspect of courtroom testimony—based upon the expectation of yes-no opinions expressed as virtual certainties—to be an impossible arena in which to pursue the truth about human behavior. As the reader will discover, this essential tension between law and psychology is a constant backdrop that shapes and informs the nature of psychological practice in the courtroom.

For better or for worse, psychologists do testify in court cases, and the focus of their testimony often pertains to the interpretation of psychological tests and assessment interviews. When are test results and psychological opinions based upon them admissible in court? What criteria do judges use in determining whether to admit psychological testimony? Psychologists who represent themselves as experts and who use tests to justify their opinions must have a firm grounding in legal issues that pertain to assessment. In this topic we examine the relevance of legal standards to testimony based upon psychological tests and evaluations. We also explore a few specialized instruments useful in forensic assessment.

The role of the psychological examiner can intersect with the legal system in a multitude of ways. The practitioner might be called upon for the following:

- Evaluation of possible malingering
- Assessment of mental state for the insanity plea
- Determination of competency to stand trial
- Assessment of personal injury
- Specialized forensic personality assessment

These are the primary applications of forensic practice, which we examine here. A variety of additional applications are surveyed in Melton, Petrila, Poythress, and Slobogin (**1998** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)**).

In addition to meeting the general guidelines for ethical practice required of any clinician, practitioners who offer expert testimony based upon psychological tests will encounter additional standards of practice unique to the U.S. jurisprudence system. We summarize major concerns regarding psychological tests and courtroom testimony here. The reader can find extended discussions of this topic in Melton et al. (**1998** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)**) and Wrightsman, Nietzel, Fortune, and Greene (**2002** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1790)**).

Each of the previously listed topics raises unique questions about the role of the psychologist in the courtroom. However, one issue is common to all forms of courtroom testimony: When is a psychologist an expert witness? We discuss this general issue before returning to specific applications of psychological evaluation that intersect with the U.S. legal system.

Standards for the Expert Witness

Just as psychologists are concerned with issues of standards and competence, so too are lawyers and judges. U.S. jurisprudence has developed various guidelines for courtroom testimony, including several general principles regarding the testimony of an **expert witness** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss107)** . These standards are found in *Federal Rules of Evidence* (**1975** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib489)**) and have been upheld by various court decisions. We can summarize the principles of expert testimony as follows:

- The witness must be a qualified expert. Not all psychologists who are asked to testify will be allowed to do so. Based on a summary of the expert's education, training, and experience, the judge decides whether the testimony of the witness is to be admitted.
- The testimony must be about a proper subject matter. In particular, the expert must present information beyond the knowledge and experience of the average juror.
- The value of the evidence in determining guilt or innocence must outweigh its prejudicial effect. For example, if the expert's testimony might confuse the issue at hand or might prejudice the members of the jury, it is generally not admissible.
- The expert's testimony should be in accordance with a generally accepted explanatory theory. In most courts, guidance on this matter is provided by *Frye v. United States*, a 1923 court case pertaining to the admissibility of expert testimony.

In *Frye v. United States*, the counsel for a murder defendant attempted to introduce the results of a systolic blood pressure deception test. The lawyer offered an expert witness to testify to the result of the deception test. It was asserted that emotionally induced activation of the sympathetic nervous system causes systolic blood pressure to rise gradually if the examinee attempts to deceive the examiner. In other words, the expert witness asserted that in the course of an interrogation about a crime, the pattern of change in systolic blood pressure could be used as a form of lie detector test. The defense counsel wanted their expert witness to testify in support of the client's innocence. Counsel for the prosecution objected, and the Court of Appeals of the District of Columbia upheld the objection, ruling:

> While courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs. (cited in Blau, 1984)

The court concluded that the systolic blood pressure deception test had not gained acceptance among physiological and psychological authorities and, therefore, refused to allow the testimony of the expert witness.

According to these guidelines, a test, inventory, or assessment technique must have been available for a fairly long period of time in order to have a history of general acceptance. For this reason, the prudent expert witness will choose well-established, extensively researched instruments as the basis for testimony, rather than relying upon recently developed tests that might not stand up to cross-examination under the constraints of *Frye v. United States*.

In the mid- to late 1990s, the standards for expert testimony were refined further, beginning with a Supreme Court decision in *Daubert v. Merrell Dow Pharmaceuticals* (1993). The Court's written opinion added extensive guidelines about factors to be considered in weighing scientific testimony in trials. Two additional court cases (*General Electric Co. v. Joiner*, 1997; *Kumho Tire Co., Ltd. v. Carmichael*, 1999) further extended the parameters of expert testimony defined by *Daubert*. Sometimes known as the *Daubert* trilogy, these three cases generated several new guidelines that trial judges may use in determining the admissibility of expert testimony (**Grove & Barden, 1999 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib659)**):

- Is the proposed theory (or technique), on which the testimony is to be based, testable?
- Has the proposed theory (or technique) been tested using valid and reliable procedures and with positive results?
- Has the theory (or technique) been subjected to peer review?
- What is the known or potential error rate of the scientific theory or technique?
- What standards controlling the technique's operation maximize its validity?
- Has the theory (or technique) been generally accepted as valid by a relevant scientific community?
- Do the expert's conclusions reasonably follow from applying the theory (or technique) to this case?

The ramifications of the *Daubert* trilogy rulings for the expert testimony of psychologists are unclear at this time. For example, it is uncertain whether testimony based upon the Rorschach Inkblot Test (discussed earlier in this text) would be admissible under these newer, more restrictive guidelines (**Grove, Barden, Garb, & Lilienfeld, 2002 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib660)**; **Ritzler, Erard, &**

Pettigrew, 2002 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1372)). What is clear at this point is that judges generally have tightened the standards for admitting expert evidence in U.S. courts (Dixon & Gill, 2002). For example, some courts have used the *Daubert* ruling as a basis for denying testimony from mental health professionals, including psychologists. In some courts, testimony about psychological evaluations of sexually abused children has been ruled inadmissible. Increasingly, courts will demand that testimony from psychologists has a strict scientific basis (**Melton et al., 1998** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)).

The Nature of Forensic Assessment

Before we turn to specific applications, it will prove helpful to explore crucial differences between forensic assessment and traditional assessment. The most general divergence is that forensic assessment is molded by the prerequisites of the legal system, whereas traditional assessment is shaped by the needs of the client and current professional standards. Although the two approaches occasionally will look the same—for example, the examiner might use the MMPI-2 in both cases—the types of information sought, the strategies for gathering it, and the manner of report writing will be noticeably different.

One major difference is the scope of evaluation in traditional and forensic assessment (**Melton et al., 1998** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)). Whereas traditional assessment usually is broadscale and provides a comprehensive picture of a client's functioning and treatment needs, forensic assessment engages a narrow focus that may not even appear to be "clinical" in nature. For example, when evaluating a forensic client for competency to stand trial, the client's symptom pattern, mental status, diagnosis, and so forth are only of tangential interest. What matters most, and what the legal system will want to know, is whether the client meets the criteria for competency or not, as discussed later in this topic. Lawyers and judges will prefer and expect a "yes" or "no" answer to the competency question, and they may regard a lengthy description of symptoms as so much drivel.

Another huge difference has to do with the client's role in the process. Whereas in traditional assessment the client voluntarily agrees to an assessment and may even help determine its scope and nature, in forensic assessment the client really has little choice in the matter, unless he or she wants to aggravate the judge who has authorized the assessment. In fact, in forensic assessment the "client" is not really the client! The psychologist is working at the behest of a judge or lawyer. Put simply, a judge, lawyer, or other court officer is usually the real client. It would be more accurate to refer to the individual undergoing the assessment as the "examinee."

Threats to validity also differ in the two settings. Although it is true that clients in traditional assessment may want to present themselves in a good light and, therefore, distort the truth when responding, this pales in comparison to the blatant faking of psychopathology (malingering) that may occur in a forensic setting. Malingering is discussed in more detail later in this topic. For now, consider the true case of an inmate evaluated by the author. He complained that he was "seeing things" and that he needed "medication" to sleep better. When asked to describe what he was "seeing," he was blandly inarticulate. During the interview, he appeared calm and collected. Furthermore, his personality test profile (MMPI-2) was a mountain range of scale elevations, a classic fake-bad profile. Beyond a doubt, he was fabricating his symptoms in hopes of receiving a prescription for antianxiety medications.

Finally, it is important to mention that the nature of the written report will differ in the two settings. In traditional assessment, the audience typically is other professionals who are familiar with jargon, diagnostic terminology, and treatment options. In forensic assessment, the audience is legal personnel who care mainly about the referral question. Melton et al. (**1998** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)) describe a number of pertinent qualities for forensic reports; namely, reports should separate facts from inferences, stay within the scope of the referral question, avoid information overkill, and minimize clinical jargon. Melton and colleagues also note that the clinician must take special care to protect the privacy rights of individuals mentioned in a report, insofar as this information most certainly will become part of the public record. Clinicians must write forensic reports with great care:

> Finally, and most important, the report and the clinician who writes it will, or at least should, receive close scrutiny during adversary negotiations or proceedings. A well-written report may obviate courtroom testimony. A poorly written report may become, in the hands of a skillful lawyer, an instrument to discredit and embarrass its author. Therefore, attention to detail and to the accuracy of information is required. (**Melton, et al., 1998** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128) , p. 523)

Woe to the forensic examiner who writes a sloppy report that becomes part of court proceedings. Not only might this unnecessarily confuse the court case, but it also could result in literally hours of ill-mannered and humiliating cross-examination of the report writer.

Evaluation of Suspected Malingering

In most settings, a psychologist safely can assume that clients will be reasonably honest about their mental and emotional state. Clients want to tell their stories and they want to get things right. At worst, they may overstate symptoms slightly so as to impress the clinician that help truly is deserved and needed. Yet outright deception and manipulation are uncommon—for the simple reason that clients rarely have incentive for these strategies.

However, the rules of clinical engagement are turned upside-down in forensic settings. The typical forensic client has much to gain from a case formulation that emphasizes illness and disability. Indeed, the context of the assessment almost guarantees that clients will seek to look "crazy" or disabled, whether by exaggeration or (more rarely) deceptive design. In the mind of the forensic client, fabrication of symptoms may serve to excuse unacceptable behavior (e.g., favoring the insanity plea), sway sentencing recommendations (e.g., against capital punishment), or gain entitlements (e.g., certification for disability). These client maneuvers clearly influence the validity of forensic assessments. Hovering in the background of every forensic assessment is this troubling question: Was the client reasonably honest and forthright?

The forensic examiner must make a judgment about the honesty of the client's self-portrayal during the evaluation. And yet while common sense dictates that the examiner should expect some degree of deception, the conclusion that a client has consciously malingered needs to be reached with caution:

> Given the significant potential for deception and the implications for the validity of their findings, mental health professionals should develop a low threshold for suspecting deceptive responding. At the same time, because the label of "malingerer" may carry considerable weight with legal decisionmakers and potentially tarnish all aspects of the person's legal position, conclusions that a person is feigning should not be reached hastily. (**Melton et al., 1998** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128))

The most common and venerable method for identifying dishonest clients is the clinical interview. However, a more objective approach should be preferred. The assessment of potential malingering with interview hinges upon the judgment of the clinician (e.g., "This client is inconsistent in his presentation of symptoms and appears eager to be sick, so I conclude that he is malingering"), which may prove erroneous. In contrast, an objective approach provides normative data, hit rates, and the like for the evaluation. Not only might this improve the accuracy of the assessment, in addition more standardized approaches should find greater acceptability in many court systems.

According to DSM-IV **malingering** is defined as:

> . . . the intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives such as avoiding military duty, avoiding work, obtaining financial compensation, evading criminal prosecution, or obtaining drugs. (**American Psychiatric Association, 1994** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib32) , p. 683)

The incidence of malingering among referred clients is hard to pin down, although thought to be significant, at least in forensic settings. Forensic practitioners estimate the occurrence to be 15 to 20 percent of their cases (**Rogers, 1986** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1388) ; **Rogers,**

Sewell, & Goldstein, 1994 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1396)). For reasons of justice and fairness, the detection of malingering with empirically validated procedures is an important obligation of forensic psychologists.

Several promising tests of malingering have emerged in recent years (**Rogers, 2008** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1391)). For reasons of space, we will focus here on three procedures that illustrate the breadth of approaches available: the Structured Interview of Reported Symptoms (SIRS), the Test of Memory Malingering (TOMM), and certain MMPI-2 indices.

Structured Interview of Reported Symptoms (SIRS)

One promising instrument is the Structured Interview of Reported Symptoms (SIRS), a 172-item interview schedule designed expressly for the evaluation of malingering (**Rogers, Bagby, & Dickens, 1992** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1382)). The approach embodied in the SIRS was based on strategies identified in the clinical literature as potentially useful for detecting malingering. Using a structured interview method, malingering is assessed on eight primary scales:

- Rare Symptoms (overreporting of infrequent symptoms)
- Symptom Combinations (real psychiatric symptoms that rarely occur together)
- Improbable or Absurd Symptoms (symptoms reveal a fantastic quality)
- Blatant Symptoms (overendorsement of obvious signs of mental disorder)
- Subtle Symptoms (overendorsement of everyday problems)
- Severity of Symptoms (symptoms portrayed with extreme, unbearable severity)
- Selectivity of Symptoms (indiscriminant endorsement of psychiatric problems)
- Reported versus Observed Symptoms (comparison of observed and reported symptoms)

In addition to the eight primary scales, five supplementary scales are used to interpret response styles. Of the 172 questions, 32 are repeated inquiries to detect inconsistency of responding. Examples of the kinds of structured interview questions include: "Do you ever feel like the fillings in your teeth can pick up radio messages?" (Rare Symptoms); "Do you have severe headaches at the same time as you have a fear of germs?" (Symptom Combinations); "Does the furniture where you live seem to get bigger or smaller from day to day?" (Improbable or Absurd Symptoms); "Do you have any serious problems with thoughts about suicide?" (Blatant Symptoms). The scale takes less than an hour to administer.

Results allow for classification of examinees as definite feigning, probable feigning, and honest. Reliability of the instrument is good, with internal-consistency reliability coefficients for subscales ranging from .66 to .92. Interrater reliability estimates are superb, ranging from .89 to 1.00.

Although the validity of the SIRS can be discussed along the familiar lines of content, criterion-related, and construct validity (and the test performs well in these domains), the real measure of its clinical utility pertains to the capacity of the test to discriminate known or suspected malingerers from psychiatric patients and normal controls. One recent study indicates that the test performs well in this capacity (**Gothard, Viglione, Meloy, & Sherman, 1996** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib624)). In a mixed sample of 125 males referred for competency evaluation (including 30 persons asked to simulate malingering, 7 individuals strongly suspected of malingering, and 88 persons for whom malingering appeared unlikely), the SIRS was overall 97.8 percent accurate in classifying participants as malingered or nonmalingered.

In a review and meta-analysis of the SIRS, Green and Rosenfeld (**2011** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib643)) found that studies published since the initial validation demonstrate higher sensitivity (correct detection of those known to be malingerers) but lower specificity (correct identification of those known to possess real psychological difficulties). In other words, genuine patient samples are more likely to be misclassified as feigning than nonclinical samples.

Another concern about the SIRS is the comparative lack of research with populations in the criminal justice system, where the instrument often is used. This population is relatively uneducated, and minorities are heavily

overrepresented, constituting 57 percent of the jail population nationwide (tables from **www.census.gov (http://www.census.gov)**). By one estimate, more than 80 percent of the urban jail population in the United States is African American (Dixon, 1995). How well does SIRS perform with incarcerated populations? In a large sample of jail inmates, McDermott and Sokolov (**2009 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1088)**) reported that 66 percent of respondents scored in the malingering range on the test. But inmates designated as malingering in their charts were no more likely than others to score in the malingering range, raising questions about the validity of the test, the accuracy of clinical judgments, or both. Based on a study of 43 individuals with intellectual disability, Weiss, Rosenfeld, and Farkas (**2011 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1747)**) recommend caution with the SIRS because a high proportion of their sample was incorrectly classified as feigning.

Test of Memory Malingering (TOMM)

The TOMM is a 50-item visual recognition test that includes two learning trials and an optional retention trial (**Tombaugh, 1997 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1653)**). The secret to the test is that it appears to be difficult, while actually it is quite easy. As a result, malingerers encounter an enticing opportunity to perform poorly, whereas others complete the task with near perfect scores. In the first learning trial, 50 line drawings are presented to the individual for 3 seconds each. Then, in the first test phase each stimulus is presented alongside three distractor drawings; the examinee is asked to pick each item shown previously. Of course, the position of each correct drawing alongside the three distractor choices is varied randomly. A second learning trial then ensues, with a second test phase. A delayed retention trial, consisting only of the test phase, is administered after a 20-minute delay. The results for the TOMM consist of the number of correct choices (out of 50) on Trial 1, Trial 2, and the Retention test.

Although there are several ways to summate and interpret TOMM test scores, the most common approach is to utilize a cutting score of 44/45 on the Trial 2 outcome. In other words, scores of 45 or higher on Trial 2 are considered normal, whereas scores of 44 or lower indicate the likelihood of malingering. This interpretive strategy emerged from several studies indicating that individuals who have no motivation to malinger—whether they are normal adults *or* patients with brain impairment—rarely score lower than 45 on the second trial. In contrast, individuals with motivation to malinger (e.g., brain-injured persons involved in litigation, or others who have something to gain from poor test performance) often score well below 45. We have summarized the score ranges for a few studies in **Table 12.3 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec5#ch12tab3)** . The three samples portrayed in this table include a large nonclinical sample of 405 adults (**Tombaugh, 1997 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1653)**), a small sample of 22 individuals with confirmed traumatic brain injury (TBI) but no motivation to malinger (i.e., no pending litigation), and a small sample of 28 individuals with mild head injury seeking compensation and therefore possessing a strong motivation to malinger (**last two samples from Haber & Fichtenberg, 2006 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib679)**). The reader will notice that 99 percent of the intact adults scored between 45 and 50, and 100 percent of the TBI patients with no motivation to malinger scored in this normal range. In contrast, a whopping 64 percent of the patients seeking compensation for head injury (and therefore having motivation to malinger) scored below the cut-off, far worse than the sample with confirmed brain damage! Recent research confirms the value of the TOMM in a variety of settings, including clinic-referred pediatric patients (**Kirk, Harris, Hutaff-Lee, and others, 2010 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib895)**), juvenile offenders (**Gast & Hart, 2010 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib570)**), and Spanish-speaking TBI populations (**Strutt, Scott, Lozano, Tieu, & Peery, 2012 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1597)**).

TABLE 12.3 Score Ranges for the Test of Memory Malingering

	Percent Obtaining TOMM Trial 2 Score of:

	50494847464540-4430-39<30
Samples:	
Intact adults no motive (N = 405)[a] (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec5#ch12fn1)	91 7 1 1 1
Definite TBI with no motive (N = 22)[b] (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec5#ch12fn2)	7314 5 5 5
Possible TBI with with motive (N = 28)[b] (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec5#ch12fn2)	25 4 0 7 0 0 18 25 21

[a]Based on data from Tombaugh (1997 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1653))

[b]Based on data from Haber & Fichtenberg (2006 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib679))

Note: TBI = traumatic brain injury. Percentages are rounded off and row totals therefore do not equal exactly 100 percent.

Assessment of Mental State for the Insanity Plea

In criminal trials the defendant may invoke a variety of defenses including entrapment, diminished capacity (e.g., from mental subnormality), automatism (e.g., from hypnotic suggestion), and the insanity plea. Whenever a special defense is invoked, an evaluation of the defendant's **mental state at the time of the offense (MSO)** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss199) is required. In some courts, a psychologist is qualified to offer opinions about the MSO of a defendant. We restrict the discussion here to the insanity plea since this is the most common doctrine that would trigger the need for an MSO evaluation.

Almost everyone is familiar with the insanity defense, but only the exceptional person understands its provisions. Technically, the insanity defense is known as **not guilty by reason of insanity (NGRI)** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss224) . Based on a few sensational and widely publicized trials such as the case of John Hinckley, who attempted to assassinate President Ronald Reagan, the lay public generally has concluded that the insanity defense is commonly employed by cynical lawyers to help dangerous clients evade legal responsibility for heinous crimes. Nothing could be further from the truth. In reality, the NGRI plea is widely respected by jurisprudence experts and is invoked in fewer than 1 in 1,000 trials (Blau, 1984). And in this tiny fraction of all criminal cases, the defense succeeds less than 1 time in 4 (**Melton et al., 1998** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)). The widespread belief that persons found NGRI "walk" away from their crimes also is inaccurate: Most receive hospital treatment that lasts several years. Recidivism rates are perhaps lower (and certainly not higher) than felons convicted of similar offenses (**Melton et al., 1998** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)). Even though outlawed in some states, the insanity defense has shown remarkable resiliency—probably because it performs a desirable role in a modern and compassionate society.

Several legal tests for insanity have had significant influence in the United States, including the M'Naughten rule, the Durham rule, the Model Penal Code rule, and the Guilty But Mentally Ill (GBMI) verdict (**Wrightsman et al., 2002** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1790)). Some jurisdictions include irresistible impulse as a supplement to the M'Naughten Rule. A few states have abolished the insanity defense altogether. We will survey the different standards briefly before commenting upon the role of psychological tests in determining legal insanity.

The **M'Naughten rule** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss206) is the oldest, stemming from an 1843 case in England. Daniel M'Naughten was plagued by paranoid delusions that the prime minister, Robert Peel, was part of a conspiracy against him. M'Naughten stalked the prime minister and, in a case of mistaken identity, shot his male secretary at No. 10, Downing Street. M'Naughten was found not guilty by reason of insanity, a verdict that touched off a national furor. In response to the furor, Queen Victoria commanded all 15 high judges of England to appear before the House of Lords and clarify the newly forged guidelines on insanity. The M'Naughten rule states:

> The jury ought to be told in all cases that every man is to be presumed to be sane, and to possess a sufficient degree of reason to be responsible for his crimes, until the contrary be proved to their satisfaction; and that to establish a defense on the grounds of insanity it must be clearly proved that, at the time of committing the act, the accused was laboring under such a defect of reason, from disease of the mind, as not to know the nature and quality of the act he was doing, or, if he did know it, that he did not know what he was doing was wrong. (cited in Wrightsman et al., 1994)

Thus, the M'Naughten rule "excuses" criminal behavior if the defendant, as a consequence of a "disease of the mind," did not know what he or she was doing (e.g., a paranoid schizophrenic who believed he or she was shooting the literal devil) or did not know that what he or she was doing was wrong (e.g., a person with mental retardation who believed that it was acceptable to shoot an obnoxious panhandler). Approximately half of the states use the M'Naughten rule.

Some jurisdictions also allow "irresistible impulse" as a supplement to the M'Naughten rule. An *irresistible impulse* is generally defined as a behavioral response that is so strong that the accused could not resist it by will or reason. But when is an impulse irresistible as opposed to simply unresisted? This has proved difficult to define. For obvious reasons, legal experts are unhappy with the notion of irresistible impulse, and its use as part of an insanity plea appears to be waning.

The **Durham rule** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss95)** was formulated in 1954 by the District of Columbia Federal Court of Appeals in *Durham v. United States.* Dissatisfied with the M'Naughten rule, Judge David Bazelon proposed a new test, known as the Durham rule, which provided for the defense of insanity if the criminal act was a "product" of mental disease or defect. The purpose of the Durham rule was to give mental health professionals a wider latitude in presenting information pertinent to the defendant's responsibility. Legal scholars hailed Durham as a great step forward, but in 1972 the rule was dropped by the circuit that had formulated it.

The Durham rule was replaced by the Model Penal Code rule proposed by the American Law Institute. Adopted in 1972, the Model Penal Code rule is as follows:

> A person is not responsible for criminal conduct if at the time of such conduct, as a result of mental disease or defect, he lacks substantial capacity either to appreciate the criminality (wrongfulness) of his conduct or to conform his conduct to the requirements of the law. (cited in **Melton et al., 1998** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)**)

The **Model Penal Code** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss208)** rule also contains provisions that prohibit the inclusion of the psychopath or antisocial personality within the insanity defense.

The Model Penal Code rule differs from the M'Naughten rule in three important ways:

- By using the term *appreciate*, it acknowledges the emotional determinants of criminal action.
- It does not require a total lack of appreciation by offenders for the nature of their conduct—only a lack of "substantial capacity."
- It includes both a cognitive element and a volitional element, making defendants' inability to control their actions an independent criterion for insanity (**Wrightsman et al., 2002** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1790)** , p. 329).

About 20 states now follow the Model Penal Code rule or slight variants of it.

A recent development in the insanity plea is the **Guilty But Mentally Ill (GBMI)** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss141)** verdict. Approximately one-fourth of the states allow juries to reach a verdict of GBMI in cases in which the defendant pleads insanity. Typically, in states that allow the GBMI verdict, the judge instructs the jury to return with one of four verdicts:

- Guilty of the crime
- Not guilty of the crime
- Not guilty by reason of insanity
- Guilty but mentally ill

The intention of the last alternative is that a defendant found GBMI should receive the same sentence as if found guilty of the crime, but he or she begins the sentence in a psychiatric hospital. After treatment is completed, the defendant then serves the remainder of the sentence in a prison.

But the intention of GBMI and its reality are two different things. Initial support for the GBMI verdict as a humane variant of the insanity plea has waned in recent years. Wrightsman et al. (**2002** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1790)**) point out that jurors express confusion when asked to make the difficult distinction between mental illness that results in insanity (GBMI) and mental illness that does not. Melton et al. (**1998** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)**) find little virtue in the verdict:

> The GBMI verdict is conceptually flawed, has significant potential for misleading the fact-finder, and does not appear to achieve its goals of reducing insanity acquittals or prolonging confinement of offenders who are mentally ill and dangerous. The one goal it may achieve is relieving the anxiety of jurors and judges who otherwise would have difficulty deciding between a guilty verdict and a verdict of not guilty by reason of insanity. It is doubtful this goal is a proper one or worth the price. (p. 215)

Empirical studies indicate that offenders found GBMI seldom receive adequate treatment. Furthermore, they may receive harsher sentences than their counterparts found merely guilty (**Callahan, McGreevy, Cirincione, & Steadman, 1992** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib253)**). In fact, some defendants found GBMI have been sentenced to death!

Now that the reader has been introduced to variants of the insanity plea, we review the role of the psychologist in determining legal insanity. An important point is that psychologists are rightfully cautious in offering an interview-based opinion as to a person's mental state at the time of a criminal offense. After all, the crime usually occurred days, weeks, months, or even years before, and the client may be unable to assist in the accurate reconstruction of events and mental states. Consequently, psychological testimony regarding legal insanity should be cautious and conservative. Reliability studies of insanity evaluations also suggest that caution is appropriate. In a review of seven studies, Melton et al. (**1998** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)**) determined that interrater agreement (as to whether a defendant was legally insane) ranged from a low of 64 percent (between prosecution and defense psychiatrists) to a high of 97 percent (between psychologists with forensic training who used structured instruments, discussed later).

In spite of controversy over the role of the psychologist in MSO determinations, some experts foresee an increased role for psychological assessment in cases involving the insanity plea. In particular, neuropsychological assessments may provide objective, valid data to help the courts decide the merits of an insanity defense. Recent court rulings affirm that neuropsychological test findings can be used to show that a defendant has impaired capability to choose right and refrain from wrong (**Blau, 1984; Heilbrun, 1992** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib731)**). Martell (**1992** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1046)**) has discussed the relevance of neuropsychological assessment to the insanity plea as defined by the Model Penal Code. The Model Penal Code defines a defendant as not guilty by reason of insanity if he or she "lacks substantial capacity" to appreciate the criminality of his or her conduct. Neuropsychological test results have a direct bearing upon this issue.

Rating scales such as the Rogers Criminal Responsibility Assessment Scales (R-CRAS) also provide a useful basis for evaluating criminal responsibility (**Rogers, 1984** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1387)** , **1986**

(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1388)). The R-CRAS is completed by the examiner immediately following a review of clinical records, police investigative reports, and the final clinical interview with the patient-defendant. The instrument consists of clear descriptive criteria for 25 items assessing both psychological and situational factors. The items are scored with respect to the time of the crime on five scales measuring these variables:

- Patient Reliability
- Organicity
- Psychopathology
- Cognitive Control
- Behavioral Control

TABLE 12.4 Sample Items from the R-CRAS

- Amnesia about the alleged crime.

  (This refers to the examiner's assessment of amnesia, not necessarily the patient's reported amnesia)

  (0) No information.
  (1) None. Remembers the entire event in considerable detail.
  (2) Slight; of doubtful significance. The patient forgets a few minor details.
  (3) Mild. Patient remembers the substance of what happened but is forgetful of many minor details.
  (4) Moderate. The patient has forgotten a major portion of the alleged crime but remembers enough details to believe it happened.
  (5) Severe. The patient is amnesic to most of the alleged crime but remembers enough details to believe it happened.
  (6) Extreme. Patient is completely amnesic to the whole alleged crime.
- Delusions at the time of the alleged crime.
  (1) No information.
  (2) Suspected delusions (e.g., supported only by questionable self-report).
  (3) Definite delusions but not actually associated with the commission of the alleged crime.
  (4) Definite delusions which contributed to, but were not the predominant force in, the commission of the alleged crime.
  (5) Definite controlling delusions on the basis of which the alleged crime was committed.

Source: Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Rogers' Criminal Responsibility Assessment Scales by Richard Rogers, Ph.D. Copyright 1984 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.

The individual items on the R-CRAS were derived from the Model Penal Code standard of insanity (**Table 12.4** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec5#ch12tab4)). Interrater reliabilities of the R-CRAS scales ranged from .48 (for a Malingering subscale) to 1.00 (for Organicity). Construct validity was established by comparing the disposition of 93 legal cases with R-CRAS data. Even though legal outcome is determined by many variables besides the psychological state of the person at the time of the crime, there was 95 percent agreement in the determination of sanity and 73 percent agreement in the determination of insanity.

Even though reviewers recognize the promise of the R-CRAS, for some a healthy skepticism still prevails. One concern is that the subscales of the instrument represent an *ordinal* level of measurement, whereas an *interval* level of quantification is implied. Another concern is that the test developers claim to "quantify areas of judgment that are logical and/or intuitive in nature" that leads to a false sense of scientific certainty (**Melton et al., 1998** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)). Certainly, the R-CRAS performs a valuable function by helping clinicians organize their thinking and evaluation. The utility of the overall decision—sane versus insane—will rest upon additional validational research (**Howell & Richards, 1989** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib791)). In support of test validity, Rogers and Sewell (**1999** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1393))

reanalyzed 413 insanity cases and found that the R-CRAS contributed substantially to the determination of criminal responsibility.

Competency to Stand Trial

The Sixth Amendment to the U.S. Constitution, passed in 1791, guarantees every accused citizen the right to an impartial, speedy, and public trial with benefit of counsel. If the defendant is unable to exercise these constitutional rights for any reason, then a proper trial cannot take place. Specifically, if the defendant has a mental defect, illness, or condition that renders him or her unable to understand the proceedings or to assist in his or her defense, the defendant would be considered incompetent to stand trial. This standard was confirmed by the U.S. Supreme Court in *Dusky v. United States* (1960) as "whether [the defendant] has sufficient present ability to consult with his lawyer with a reasonable degree of rational understanding—and whether he has a rational as well as factual understanding of the proceedings against him." In practice, **competency to stand trial** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss55)** refers to four elements and distinctions (**Melton et al., 1998** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)**):

- The defendant's capacity to understand the criminal process, including the role of the participants in that process
- The defendant's ability to function in that process, primarily through consulting with counsel in the preparation of a defense
- The defendant's capacity, as opposed to willingness, to relate to counsel and understand the proceedings
- The defendant's reasonable degree of understanding, as opposed to perfect or complete understanding

Most U.S. courts follow this standard, which emphasizes *current* functioning of the accused.

The presiding judge may request a psychological or psychiatric evaluation to assist in determining a defendant's competency to stand trial. One recent report indicates that more than 25,000 evaluations of competency to stand trial are performed in the United States each year (**McDonald, Nussbaum, & Bagby, 1992** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1089)**). It is important to emphasize that psychologists, psychiatrists, and other mental health professionals merely assist in a competency hearing by presenting expert opinions. Only the judge has the power to make a competency determination. Although there is no standard format for a competency determination, most judges request that the psychologist consider most or all of the 11 factors cited in **Table 12.5** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec5#ch12tab5)**).

Incompetency to stand trial is entirely separate from legal insanity; these two issues are judged by completely different standards. *Legal insanity* pertains to the moment of the criminal act, whereas *incompetency* implies a current, ongoing condition. Furthermore, incompetency is not synonymous with mental illness, although the two may occur together. In the event that the judge rules the defendant incompetent, the trial is postponed, usually for a period of six months or so. In some cases, persons found incompetent are placed in a mental institution for treatment to restore their competency so that a trial can be held later. Individuals charged with less-serious crimes may receive outpatient treatment.

In addition to information obtained from the clinical interview, psychological test results are important components of a competency evaluation. For example, a low IQ may constitute evidence of incompetence in the eyes of the court. Although there are no firm guidelines, most courts rule that persons with significant intellectual deficits—say, an IQ in the range of moderate mental retardation or lower—are incompetent to stand trial. Likewise, a pattern of test results indicating severe neuropsychological deficit may warrant a finding of legal incompetence, even if the client's IQ is in the normal range. For example, a defendant with severe stroke-induced deficits in language comprehension may be found incompetent to stand trial.

TABLE 12.5 Factors Considered in Determining Competency to Stand Trial

- Defendant's appreciation of the charges
- Defendant's appreciation of the nature and range of penalties
- Defendant's understanding of the adversary nature of the legal process
- Defendant's capacity to disclose to attorney pertinent facts surrounding the alleged offense
- Defendant's ability to relate to attorney
- Defendant's ability to assist attorney in planning defense
- Defendant's capacity to realistically challenge prosecution witnesses
- Defendant's ability to manifest appropriate courtroom behavior
- Defendant's capacity to testify relevantly
- Defendant's motivation to help himself in the legal process
- Defendant's capacity to cope with the stress of incarceration prior to trial

---

Source: Florida Rules of Criminal Procedure, cited in Wrightsman, L. S., Nietzel, M. T., & Fortune, W. H. (1994). *Psychology and the legal system* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Several formalized screening tests and procedures are available to assist in competency evaluation. Rogers and Johansson-Love (**2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1392)**) provide an outstanding introduction to evidence-based practice in evaluating competency to stand trial. We focus our attention here on the MacArthur Competence Assessment Tool—Criminal Adjudication (MacCAT-CA), one of the most promising and psychometrically sound of the many tests developed for this purpose (**Hoge, Bonnie, Poythress, & Monahan, 1999** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib761)**).

The MacCAT-CA consists of 22 items grouped into three subscales of psycholegal abilities: Understanding, Reasoning, and Appreciation. The examiner begins the test by reading a hypothetical short story to the defendant about two men who get into a fight (one of them is later charged). The first subscale (8 items) assesses the defendant's ability to understand the legal system with questions like "What is the job of the defendant's lawyer?" These questions are scored 0, 1, or 2, based on degree of understanding. The second subscale (8 items) evaluates the defendant's ability to reason in regard to the hypothetical story, and to evaluate legal options for the hypothetical defendant. The third subscale (6 items) departs from the imaginary scenario and assesses the defendant's capacity to understand his or her legal situation. The questions explore the defendant's appraisal of how he or she is likely to function and to be treated during the course of the trial.

The psychometric properties of the MacCAT-CA were evaluated in a study of 729 felony defendants (**Poythress, Monahan, Bonnie, Otto, & Hoge, 2002** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1321)**). The researchers found good internal consistency for the three subscales, with coefficient alphas of .81 to .88. The construct validity of the instrument is well supported by confirmatory factor analyses that yield the three factors posited by the test developers (**Zapf, Skeem, & Golding, 2005** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1800)**). In general, the MacCAT-CA and similar instruments are a useful beginning to a competency evaluation but should not be the sole method of assessment. Most forensic experts emphasize the complexities of the legal process and the need to use competency screening instruments sparingly, and mainly in complex cases, as an adjunct to the clinical interview. The MacCAT-CA and similar approaches prove less helpful when clients put forth little effort, demonstrate cognitive impairment, or come from diverse cultural backgrounds (**Pinals, Tillbrook, & Mumley, 2006** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1301)**). Additional competency screening tests are reviewed by Zapf and Roesch (**2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1799)**).

A serious concern in competency evaluations is whether the client is malingering. After all, delaying a trial date for a long time provides a strong motive to appear incompetent. Clinicians have a variety of methods and tests (described previously) for identifying clients who might be malingering. Even so, the process of competency evaluation is not foolproof, as indicated by such high-profile cases as the Connecticut man who avoided prosecution for murder (Associated Press, June 30, 1998). This individual had allegedly murdered his former

girlfriend and her current boyfriend with a handgun, then shot himself in the head. He suffered brain damage and partial paralysis and was declared incompetent to stand trial by *four* psychiatrists in four separate hearings. They argued that he was incapable of communicating effectively with his lawyer. A court order that he undergo yearly competency evaluations was overturned, dropping him through the cracks and leaving him a free man. Nine years later he was found attending college as a pre-med student with a 3.3 grade point average. Examples like this are reason for humility and caution when psychologists approach competency evaluations.

Personal Injury and Related Testimony

Personal injury as from an automobile accident is often a source of litigation for monetary compensation. In **personal injury** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss238)** lawsuits, attorneys may hire psychologists to testify as to the lifelong consequences of traumatic stress or acquired brain damage. For example, a clinical neuropsychologist might administer a comprehensive test battery (see **Chapter 10 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch10#ch10)** , Neuropsychological Assessment and Screening) and then testify as to the long-term functional implications of known brain damage.

In general, a consulting psychologist who testifies in court will encounter extremely high practice standards. We have already mentioned the *Frye* standard, which provides that testimony must be based upon tests and procedures that have "gained general acceptance" in the field. Thus, a test or procedure that is relevant or useful in everyday clinical practice—but which is not widely accepted in the field—might be greeted with skepticism in the courts. A judge may even rule that testimony is inadmissible if it is based on tests or procedures with flimsy validation. Worse yet, the judge may allow such testimony, which opens the expert witness to criticism and ridicule by opposing attorneys. With these concerns in mind, Heilbrun (**1992 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib731)**) has published guidelines for the practice of forensic assessment. These include the use of well documented tests with reliability of .80 or higher, interpretation with actuarial formulas where available, and evaluation for malingering, defensiveness, and other reasons to discount the test data.

Increasingly, courts have been willing to compensate mental injuries in addition to physical injuries. The damage is variously referred to as "psychic trauma" or "emotional distress" or "emotional harm." The evaluation of emotional injury will rely somewhat on psychological test results (especially personality tests), but the assessment requires great clinical skill including "a longitudinal history of the impairment, its treatment, and attempts at rehabilitation, including the claimant's motivation to recover" (**Melton et al., 1998 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)** , p. 381). We see once again that the question of malingering haunts most forms of forensic assessment.

Specialized Personality Assessment in Forensic Settings

On occasion, psychologists are asked to provide specialized forms of personality assessment in forensic settings. For example, a prison psychologist might evaluate an inmate for antisocial tendencies, or a forensic psychologist might assess a treated pedophile for sexual interest in young children. The range of tools and techniques useful for specialized assessment is broad. We cover only one specialized approach here.

In prison settings, examiners have a special interest in determining whether inmates possess the traits of psychopathic personality. Similar to antisocial personality as described in *DSM-IV* (**APA, 2000 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib33)**), the concept of psychopathic personality has a long and rich history that dates back to Emil Kraepelin (1856–1926), the father of diagnostic psychiatry. But it was the psychiatrist Cleckley (**1941 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib312)**) who first provided a detailed description of the psychopath in his pathbreaking book, *The Mask of Sanity*. Based upon extensive clinical work with individuals who he labeled psychopaths, Cleckley identified a number of personality traits and behavioral signs displayed by these individuals. The key qualities appear to be a lack of remorse or shame in a charming individual who uses other people and whose life lacks any goal or direction. Good at lying, the psychopath also shows poor judgment and is considered incapable of love.

Although his description is antiquated in places, contemporary researchers continue to find descriptive and predictive value in Cleckley's conception of the psychopath. In fact, one researcher has developed a highly respected and widely used assessment tool based closely on this original formulation of psychopathic personality.

The Psychopathy Checklist-Revised (**PCL-R; Hare, 2003** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib699)** ; **Hare & Neumann, 2006** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib702)**) consists of a 20-item rating scale carefully designed to assess the qualities of psychopathic personality in a quantitative and empirical fashion. Prior to filling out the rating scale, the examiner conducts a lengthy semistructured interview (90 to 120 minutes) with the client. The interview concerns the Cleckley-based traits of psychopathy, as slightly revised and expanded by Hare and colleagues (**Hare, Harpur, Hakstian, and others, 1990** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib701)**). Each item reflects a particular symptom, such as glibness and superficial charm, grandiose sense of self-worth, pathological lying, lack of remorse or guilt, or failure to accept responsibility. Items are rated on a 3-point scale (0 = doesn't apply, 1 = applies somewhat, 2 = definitely applies). The rating is based on lifetime functioning rather than the present state, which explains why a long interview is essential for correct use of the scale. The item scores are then summed to yield a total score (range of 0 to 40) that reflects the extent to which the individual resembles the prototypical psychopath. Two factor scores also can be derived, each based on eight or nine items. Factor 1 reflects the selfish, callous, and remorseless use of others, whereas factor 2 indicates a chronically unstable and antisocial lifestyle.

A substantial body of research indicates that the PCL-R possesses strong reliability and validity. For example, interrater reliability is typically in the .90s, test-retest coefficients approach .90, and internal consistency coefficients are in the mid-to high .80s (**Schroeder, Schroeder, & Hare, 1983** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1473)**). The predictive validity of the instrument is bolstered by the capacity of PCL-R scores to predict a variety of antisocial behaviors, including violent recidivism following prison release, poor response to correctional treatment programs, and disorderly behavior while in prison (**Hare, 1996** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib698)** ; **Sreenivasan, Walker, Weinberger, Kirkish, & Garrick, 2008** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1543)**). For example, in one study inmates with high PCL-R scores were twice as likely to engage in fights and more than three times as likely to be belligerent than other inmates (**Hare & McPherson, 1984** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib700)**). In another study of a therapeutic community treatment program for adult male offenders, psychopaths showed less clinical improvement, demonstrated lower levels of motivation, and were discharged from the program earlier than nonpsychopaths. The differences were not small: The psychopathic lawbreakers stayed in the treatment program less than *half* as long as the other offenders (**Ogloff, Wong, & Greenwood, 1990** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1251)**). Clearly, psychopathic personality is a useful concept, and the PCL-R is a practical measure of the construct.

One recent study does raise concerns about examiner differences in the scoring of the PCL-R in field settings (**Boccaccini, Turner, & Murrie, 2008** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib172)**). Although the manual for the instrument does provide specific guidelines for each item, judgment is needed to differentiate scores of 0, 1, or 2. It is possible for examiners to differ in their scoring tendencies for a variety of reasons, including the evaluators' readiness to seek outside information, diligence in following the required protocol, response bias in using higher or lower ends of the scales, and drift over time in observance of scoring rules. Even the characteristics of the examiner (e.g., warm versus cold demeanor) can cause item scores to shift up or down.

Boccaccini et al. (**2008** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib172)**) examined the PCL-R total scores for 20 different Texas state-contracted evaluators who were hired to screen 321 referrals for civil commitment as sexually violent predators. The evaluators encountered their referrals in a more or less quasi-random manner, so it is reasonable to expect that the average PCL-R scores across examiners should be reasonable similar. This proved not to be the case in dramatic fashion. Restricting the comparison to examiners

completing at least 20 evaluations (to provide constancy of results), the researchers found sizable disparities in average PCL-R scores: the means varied from a low of 17.5 (SD of 8.8) to a high of 27.1 (SD of 6.1). These are large inequities on the PCL-R, which has a maximum possible score of 40. In this study, examiner differences account for a large degree of variability in the PCL-R total scores. There may be a need to improve the field reliability of assessment for this forensic instrument. Perhaps some form of training program is needed to certify individuals in its use.

TOPIC 12B Computerized Assessment and the Future of Testing

12.6 Computers in Testing: Overview and History
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec6#ch12lev1sec6)

12.7 Computer-Based Test Interpretation: Current Status
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec7#ch12lev1sec7)

12.8 Interactive Video, Virtual Reality, and Smartphones
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec8#ch12lev1sec8)

12.9 Evaluation of Computer-Based Test Interpretation
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec9#ch12lev1sec9)

12.10 Computerized Adaptive Testing
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec10#ch12lev1sec10)

12.11 The Future of Testing
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec11#ch12lev1sec11)

Computers are now used in virtually every aspect of assessment, including the administration, scoring, and interpretation of many tests. In fact, for many instruments it is now possible for the practitioner to seat the client in front of a computer with instructions that consist of "Please follow the instructions." Minutes later, the practitioner receives a lengthy narrative report, consisting not only of summary scores but also a lengthy, sophisticated interpretive report. Although the use of computers in testing is manifestly a positive development, it also raises a number of troubling questions. In this topic, current applications of the computer in psychological assessment are surveyed, and the professional and social issues raised by this practice are discussed. The topic closes with thoughts on the future of testing—which will be forged in large measure by increasingly sophisticated applications of computer technology. We begin with an overview and history of the computer in testing.

12.6 COMPUTERS IN TESTING: OVERVIEW AND HISTORY

Introduction to Computer-Aided Assessment

In many counseling centers it is possible for a client to make an appointment with a microcomputer to explore career options. Other than a brief interaction with the receptionist to schedule time at the computer, the client need not interact with any other human being during the entire assessment process. The exact scenario will differ from one setting to the next but might resemble the following. Instructions on the computer screen encourage the user to press any key. The computer then prompts the client to answer a series of questions about activities and interests by pressing designated numeric keys. After completion of the inventory, the computer calculates raw scores for a long list of occupational scales and makes appropriate statistical transformations. Next, a brief report appears on the screen. The report provides a list of careers that best fit the interests of the client. A hard copy is also printed for later review. Presumably, the client is better informed about compatible career options and, therefore, more likely to choose a satisfying line of work. This scenario is a simple example of computer-assisted psychological assessment (CAPA), a recent development hailed by many psychologists but criticized by others.

It is common knowledge that computers are now used widely in psychological testing. However, the breadth of these applications might surprise the reader. In addition to straightforward applications such as presenting test questions, scoring test data, and printing test results (as described earlier), computers can be used to (1) design individualized tests based upon real-time feedback during testing, (2) interpret test results according to complex decision rules, (3) write lengthy and detailed narrative reports, and (4) present test stimuli in engaging and realistic formats, including high-definition video and virtual reality. We touch upon all of these topics in our review. The umbrella term **computer-assisted psychological assessment** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss57)** (CAPA) refers to the entire range of computer applications in psychological assessment. CAPA holds great promise to the practice of psychology but also presents a variety of practical and ethical problems that demand careful and thoughtful consideration. A brief history of CAPA is a good backdrop to the discussion of practical and ethical concerns (**Table 12.6** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec6#ch12tab6)**)

TABLE 12.6 Historical Landmarks in CAPA

1946	Hankes develops an analog computer to score the SVIB (**Moreland, 1992** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1173)**).
1954	Meehl's (**1954** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1116)**) book *Clinical versus Statistical Prediction* sets the stage for automated test interpretation.
1962	Optical scanner and digital computer are used to score SVIB and MMPI and also to print profiles (**Moreland, 1992** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1173)**).
1962	First computer-based test interpretation system is developed for the MMPI at the Mayo Clinic (**Swenson et al., 1965** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1606)**).
1964	Piotrowski publishes a system for computer-based interpretation of the Rorschach (**Piotrowski, 1964** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1305)**).
1960s	Computer-based interpretive systems for the MMPI proliferate; Fowler, Finney, and Caldwell develop popular systems (**Fowler, 1985** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib526)**).
1971	A mainframe computer with terminals is used to automate the entire assessment process for psychiatric inpatients at the VA Hospital in Salt Lake City, Utah (**Klingler, Miller, Johnson, & Williams, 1977** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib903)**).
1975	First automated interpretation of a neuropsychological test battery (**Adams & Heaton, 1985** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib09)**).

1979 Lachar publishes an actuarially based interpretive system for the Personality Inventory for Children (**Lachar & Gdowski, 1979** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib941)**).

1970s Computerized adaptive testing (CAT) is introduced; CAT allows for flexible, individualized test batteries which produce a given level of measurement accuracy with the fewest possible test items (Weiss, 1982).

1985 A special series on computerized psychological assessment appears in the *Journal of Consulting and Clinical Psychology* (**Butcher, 1985** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib238)**).

1986 American Psychological Association publishes *Guidelines for Computer-Based Tests and Interpretations.*

1987 Publication of the first resource book titled *Computerized Psychological Assessment: A Practitioner's Guide* (**Butcher, 1987** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib242)**).

1994 Introduction of multimedia assessment batteries; for example, at IBM, a multimedia test is used to assess the real-life problem-solving skills of prospective employees (*APA Monitor*, June 1994).

1997 Educational Testing Service and other testing giants move to computerized testing for major admissions tests such as the Graduate Management Admission Test (GMAT) and Graduate Record Examinations (GRE).

12.7 COMPUTER-BASED TEST INTERPRETATION: CURRENT STATUS

Computer-based test interpretation
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss58), or CBTI, refers to test interpretation and report writing by computer. Every major test publisher now offers computer-based test interpretations. These services are available by mail-in, online computer with modem, or on-site microcomputer package. Moreover, the market for computer-based testing and report writing is so lucrative that we can anticipate massive growth in this field for many years to come. Butcher (**1987**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib242), App. A) listed 169 vendors as of 1986. Conoley, Plake, and Kemmerer (1991) note that the number of computerized psychological test interpretations had increased to more than 400 by 1990. New computerized test systems are reported virtually every month in trade magazines and newspapers (e.g., *APA Monitor*). Computer-based test interpretation is here to stay.

In this section we will provide an overview of the types of computer-based test interpretations currently available. A comprehensive review of products could easily span several volumes, so the reader will have to settle for a discussion of diverse and representative examples of CBTI. We will examine four approaches to CBTI: scoring reports, descriptive reports, actuarial reports, and computer-assisted clinical reports (**Moreland, 1992**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1173)).

Scoring Reports

Scoring reports consist of scores and/or profiles. In addition, a scoring report may include statistical significance tests and confidence intervals plotted for the test scores. By definition, scoring reports do not include narrative text or explanation of scores. Moreland (**1992**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1173)) discusses the appeal of scoring reports:

> These kinds of data make it possible to identify especially meaningful scores and meaningful differences among scores at a glance. They should also increase a user's confidence that those scores are in fact important. Statistical significance tests are undoubtedly superior to "clinical rules of thumb" when it comes to accurate interpretation of test scores. And who has time to hand calculate confidence intervals—especially for tests with dozens of scales?

An example of a scoring report for the Jackson Vocational Interest Survey (**Jackson, 1991**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib819)) is shown in **Figure 12.1**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec7#ch12fig1). The reader will notice that a great deal of information is presented in an efficient, condensed manner. This is typical of scoring reports. In a single page, this hypothetical respondent would learn that his interests are highly similar to majors in liberal arts, education, and business. In terms of occupational fit, he also learns that he is highly compatible with counselors, teachers, lawyers, administrators, and other professions with an emphasis upon human relations.

Descriptive Reports

A descriptive report goes one step further than a scoring report by providing brief scale-by-scale interpretation of test results. Descriptive reports are especially useful when test findings are conveyed to mental health professionals who have little knowledge of the test in question. For example, most clinical psychologists know that a high score on the MMPI Psychasthenia scale signifies worry and dissatisfaction with social relationships—but other mental health practitioners may not have a clue as to the meaning of an elevation on this scale. A descriptive report can convey invaluable information in a half page or less. A variety of descriptive reports have been developed over the years. We have provided a generic example in **Figure 12.2**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec7#ch12fig2). The reader will notice that the 20-year-old male patient is described as shy, sensitive, worried, and severely depressed. Referral of this medical patient to a psychologist or psychiatrist clearly is warranted. This report is a model of simplicity and

clarity. By comparison, most contemporary computer-based descriptive reports provide excessive detail. Typically, the clinician must wade through several pages of narrative to extract essential features about the client.

Actuarial Reports: Clinical versus Actuarial Prediction

The actuarial approach to computer-based test interpretation is based upon the empirical determination of relationships between test results and the criteria of interest. The nature of this approach is best understood in the context of the longstanding debate on clinical versus actuarial prediction. A brief detour is needed here to introduce relevant concepts and issues before discussing actuarial reports.

Many computer-based test interpretations make predictions about the test taker. These predictions are often disguised in the language of classification or diagnosis, but they are predictions nonetheless. For example, when a computer-based neuropsychological test report tentatively classifies a client as having brain damage, this is actually an implicit prediction that can be confirmed or dis-confirmed by external criteria such as brain scans and neurological consultation. Likewise, when a computer-based MMPI-2 report provides a tentative *DSM-IV* diagnosis of a clinic referral, this is also a prediction that can be validated or invalidated by external criteria such as intensive clinical interview. A final example: When a computer-based CPI screening report for police candidates warns that an applicant will make a poor adjustment in law enforcement, this is also a prediction that could be proved correct or incorrect by an inspection of personnel records at a later date.

The use of computers for test-based prediction highlights an essential distinction known as clinical versus actuarial judgment (**Dawes, Faust, & Meehl, 1989**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib397) ; **Garb, 1994**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib561) ; **Meehl, 1954**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1116) , **1965**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1117) , **1986**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1118)). In **clinical judgment**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss49) , the decision maker processes information in his or her head to diagnose, classify, or predict behavior. An example: A clinical psychologist uses experience, intuition, and textbook knowledge to determine whether an MMPI profile indicates psychosis. Psychosis is a broad category that includes serious mental disorders often characterized by hallucinations, delusions, and disordered thinking. Thus, a clinician's prediction of psychosis (or lack thereof) can be validated against external criteria such as detailed interview.

Similarity to College Students

JVIS profiles from over 10,000 university students who were enrolled in more than 150 different major fields, ranging from accounting to zoology, have been collected and analyzed. That analysis indicated that the major fields could be classed into 17 broad academic clusters. Each cluster is based on data from both males and females and represents a set of educational majors that shared a similar pattern of JVIS scores.

The chart below ranks the similarity of your JVIS Basic Interest profile to each of the student clusters. A high score indicates that your pattern of interests is similar to students in the fields of concentration defining the cluster, while a low score indicates dissimilarity. These scores indicate your probable interest and satisfaction with these academic clusters. These scores do not tell you whether or not you will be successful in any particular field.

Score	Similarity	University Major Cluster
+0.62	Very Similar	Environmental Resource Management
+0.55	Similar	Health, Physical Education and Recreation
+0.39	Moderately Similar	Agribusiness and Economics
+0.37	Moderately Similar	Art and Architecture
+0.30	Moderately Similar	Food Science
+0.12	Neutral	Engineering
+0.03	Neutral	Science
−0.03	Neutral	Computer Science
−0.08	Neutral	Performing Arts
−0.12	Neutral	Social Service
−.12	Neutral	Health Services and Science
−.19	Neutral	Mathematical Sciences
−.25	Neutral	Business
−.25	Dissimilar	Communication Arts
−.30	Dissimilar	Behavioral Science
−.32	Dissimilar	Education
−.54	Dissimilar	Social Science, Law and Politics

Your JVIS profile is most similar to college students whose academic areas of specialization are in the three clusters listed below. Sample majors for each of these three areas are also listed.

University Major Cluster	Sample Majors
Environmental Resource Management	Wildlife Technology, Recreation and Parks, Environmental Resource Management, Agricultural Business Management, Agriculture, Forest Science and Technology, Horticulture.
Health, Physical Education and Recreation	Health and Physical Education, Recreation and Parks.
Agribusiness and Economics	Agricultural Economics and Rural Sociology, Agricultural Business Management, Food Service and Housing Administration.

FIGURE 12.1 A Scoring Report from the online version of the Jackson Vocational Interest Blank

Note: The full report consists of an 11-page printout.

Source: Reprinted with permission from ***JVIS.com (http://JVIS.com)***.

Brief Descriptive Report for John Sample, 35-year-old married man with 20 years of education

Scale:	VRIN	TRIN	L	F	K					
Score:	47	50	48	53	46					
Scale:	Hs	D	Hy	Pd	Mf	Pa	Pt	Sc	Ma	Si
Score:	56	82	60	59	48	67	78	52	40	76

Major Features:

Validity: Valid profile from a cooperative patient with openness and lack of exaggeration

D	82	Depressed and pessimistic, lacks energy and concentration, thinks about self-harm
Pt	78	Worried and full of turmoil, feels anxious and tense, reports physical complaints
Si	76	Introverted and shy, uncomfortable in large social events, prefers a few close friends
Pa	67	Sensitive and moralistic, feels victimized and suspicious, tends to blame others
Hy	60	Mildly demanding and self-centered, lacks insight, somewhat attention seeking
Pd	59	Mildly rebellious and impulsive, restless, independent, energetic, and assertive
Hs	56	No special interpretation, average level of bodily and health concerns
Sc	52	No special interpretation, absence of unusual beliefs or strange behaviors
Mf	48	No special interpretation, mix of esthetic pursuits and interests in the outdoors
Ma	40	Lacking in energy, probably feels fatigued, possible depressive symptoms

FIGURE 12.2 Generic Example of an MMPI-2 Brief Descriptive Report

In **actuarial judgment** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss03)** , an empirically derived formula is used to diagnose, classify, or predict behavior. An example: A clinical psychologist merely plugs scale scores into a research-based formula to determine whether an MMPI profile indicates psychosis. The actuarial prediction, too, can be validated against appropriate external criteria.

The essence of actuarial judgment is the careful development and subsequent use of an empirically based formula for diagnosis, classification, or prediction of behavior. A common type of actuarial formula is the regression equation in which subtest scores are combined in a weighted linear sum to predict a relevant criterion. But other statistical approaches may work well for decision making, too, including simple cutoff scores and rule-based flow charts. Of course, statistical rules lend themselves to computer implementation, so it is fitting to discuss clinical versus actuarial judgment in this section on computer-based test interpretation.

Although computers facilitate the use of the actuarial method, we need to emphasize that "actuarial" and "computerized" are not synonymous. To be truly actuarial, test interpretations must be automatic (prespecified or routinized) and based on empirically established relations **(Dawes, Faust, & Meehl, 1989 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib397)**). If a computer program incorporates such automatic, empirically based decision-making rules, then it is making an actuarial prediction. Conversely, if a computer program embodies the thinking and judgment of a clinician—no matter how wise that person is—then it is making a clinical prediction.

Meehl (**1954 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1116)**) was the first to introduce the issue of clinical versus actuarial judgment to a broad range of social scientists. He stated the issue with pure simplicity: "When shall we use our heads instead of the formula?" Consider the practical problem of distinguishing between neurosis and psychosis on the basis of MMPI results. *Neurosis* is an outdated (but still used) diagnostic term that refers to a milder form of mental disorder in which symptoms of anxiety or dysphoria predominate. As noted previously, psychosis is a more serious form of mental disorder that may include hallucinations, delusions, and disordered thinking. The differential diagnosis between these two broad classes of mental disorder is important. Persons with neurosis often respond well to individual psychotherapy, whereas a patient with psychosis may need powerful antipsychotic medications that produce adverse side effects. Which is superior for MMPI-based diagnostic decision making, the head of the well-trained psychologist or an appropriate formula based upon prior research? We return to this issue later.

Meehl (**1954 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1116)**) specified two conditions for a fair comparison of these contrasting approaches to decision making. First, both methods should base judgments on the same data. For example, in comparing the experienced clinician against an actuarial

equation, both approaches should prognosticate from the same pool of MMPI profiles and only those profiles. Second, we must avoid conditions that can artificially inflate the accuracy of the actuarial approach. For example, the actuarial equation should be derived on an initial sample, prior to the comparison with clinical decision making on a new sample of MMPI profiles. Otherwise, the actuarial decision rules will capitalize on chance relations among variables and produce a spuriously high rate of correct decisions.

When the conditions are met for a fair test of clinical versus actuarial decision making, the latter method is superior in the vast majority of cases. The actuarial approach is clearly better for the task cited previously—differential diagnosis of neurosis or psychosis from the MMPI. L. R. Goldberg (**1965** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib603)**) determined that a simple linear sum of selected MMPI scale scores resulted in 70 percent correct classifications, whereas Ph.D. psychologists averaged only 62 percent, with the single best psychologist achieving 67 percent correct decisions. The decision rule that defeated all human contenders was: if the *T*-score sum on L + Pa + Sc − Hy − Pt exceeds 44, diagnose psychosis; otherwise, diagnose neurosis.[3] **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec7#ch12fn03)**

Dawes, Faust, and Meehl (**1989** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib397)**) cited nearly 100 comparative studies in the social sciences. In almost every case, the actuarial method equaled or surpassed the clinical method, sometimes substantially. The research by Leli and Filskov (**1984** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib968)**) is typical in this regard. They studied the diagnosis of progressive brain dysfunction based upon neuropsychological testing. An actuarial decision rule derived from one set of cases was applied to a new sample with 83 percent correct identification. Working from precisely the same test data, groups of inexperienced and experienced clinicians correctly identified only 63 percent and 58 percent of the new cases, respectively. The reader will notice the disturbing and embarrassing fact that experience did not improve hit rates for this clinical decision-making task.

A study by McMillan, Hastings, and Coldwell (**2004** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1109)**) also illustrates the value of simple actuarial methods for predicting clinical outcomes. Their investigation involved 124 residents of a forensic intellectual disability hospital in England. In this setting, violence is not a rare occurrence, and predicting who might be violent (and therefore require greater attention) is of utmost importance. The researchers compared two approaches to prediction of hospital violence in the next six months: (1) the actuarial approach was merely to tally the number of documented episodes in the prior six months and use this information as the index of risk; and (2) the clinical approach was to use the judgment of the clinical team (psychiatrist, psychologist, nursing staff, and attendants) on a 9-point risk scale (0 = 'no risk' and 8 = 'very high risk'). Briefly, the actuarial approach proved slightly but not significantly superior to the clinical approach. Both approaches revealed strong predictive validity. The results substantiate the common adage that "the best predictor of future behavior is past behavior."

A recent meta-analysis of 136 studies by Grove, Zald, Lebow, Snitz, and Nelson (**2000** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib661)**) provides additional support for the superiority of actuarial prediction over clinical prediction. These researchers analyzed diverse studies in the fields of medicine, education, and clinical psychology in which practitioners predicted such outcomes as academic performance, job success, medical diagnosis, psychiatric diagnosis, criminal recidivism, and suicide. In each study, the clinical predictions of the practitioners (physicians, professors, and psychologists) were compared to the actuarial predictions derived from empirically based statistical formulas. Although the researchers found a few scattered instances in which the clinical method was notably more accurate than the statistical method, on the whole, their survey confirmed prior findings on this topic. The authors conclude:

> Even though outlier studies can be found, we identified no systematic exceptions to the general superiority (or at least material equivalence) of mechanical prediction. It holds in general medicine, in mental health, in personality, and in education and training settings. It holds for medically trained judges and for psychologists. It holds for inexperienced and seasoned judges. (p. 25)

Perhaps the most disturbing conclusion of these researchers was that the availability of clinical interview actually *detracted* from the accuracy of practitioner predictions in the diverse fields studied. Compared to the empirically based statistical predictions, the clinical predictions were outperformed by an even *greater* margin when information from clinical interview was available to the practitioners. The reasons for this are unclear but likely include the susceptibility of humans to certain cognitive biases (e.g., paying too much attention to vivid interview information). Also, clinicians typically do not receive adequate feedback as to the accuracy of their judgments and, hence, have no basis for correcting maladaptive predictions.

The lesson to be learned from this literature is that computerized narrative test reports should incorporate actuarial methods, when possible. For example, computer-generated reports should use existing actuarial formulas to determine the likelihood of various psychiatric diagnoses, rather than relying upon the programmed logic of a master clinician. Unfortunately, as the reader will discover in the following, most computerized narrative test reports are clinically based—which raises concerns about their validity.

Actuarial Interpretation: Sample Approach

The developers of the Personality Inventory for Children (PIC) produced an exemplary system for computer-based actuarial test interpretation, which we will describe for illustrative purposes. The reader will recall from a previous chapter that the PIC, now updated as the PIC-2, is a true-false inventory that the parent or caregiver completes with respect to the child's behavior. Based upon these responses, a profile of *T* scores (mean of 50, SD of 10) is produced for four validity scales (e.g., Defensiveness), 12 clinical scales (e.g., Delinquency), and four factor scales (e.g., Social Incompetence). In total, *T* scores are reported for 20 scales on the PIC. Of course, higher *T* scores indicate a greater likelihood of psychopathology.

Actuarial interpretation of the PIC rests upon the empirically derived correlations between individual scales and important nontest criteria. Research subjects for the Lachar and Gdowski (**1979 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib941)**) study consisted of 431 children referred to a busy teaching clinic. As part of the evaluation process for each child, the staff members, parents, and teachers completed a comprehensive questionnaire, which listed 322 descriptive statements concerning behavior and other variables. In addition, parents or caretakers filled out the PIC.

In the first phase of the actuarial study, the 322 descriptive statements were correlated with the 20 PIC scales to identify significant scale correlates. In the second phase, the significant correlates were analyzed further to determine the relationship between descriptive statements and *T*-score ranges on the PIC scales. The outcome of this prodigious effort was a series of actuarial tables not unlike the tables used by insurance companies to predict the likelihood of illness, death, accidents, and the like, based upon population demographics such as age, sex, and residence. Some examples of actuarial correlates of the Delinquency, or DLQ, Scale are depicted in **Table 12.7 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec7#ch12tab7)** .

Actuarial tables capture a wealth of information useful in clinical practice. Consider two hypothetical 12-year-old children, Jimmy and Johnny, each referred to a clinician with the same presenting problem: school underachievement. As part of the intake procedure, the clinician asks each mother to fill out the PIC. Suppose that the Delinquency, or DLQ, Scale score for Jimmy is highly elevated at a *T* score of 114, whereas Johnny obtains an average range *T* score of 54. Based upon these scores, the clinician would know the likelihood—listed here as percentages—that certain behavioral descriptions apply to each child:

TABLE 12.7 Occurrence Rates for Actuarial Descriptors of the PIC Delinquency Scale

		T-Score Ranges							
Descriptor[(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec7#ch12fn9)]	*Base Rate*[*]	*30–59*	*60–69*	*70–79*	*80–89*	*90–99*	*100–109*	*110–119*	*>120*
Refuses to go to bed	30	18	26	23	33	36	33	42	38

Lies	62	44	36	48	73	71	79	90	91
Uses drugs	12	0	2	6	7	11	18	32	53
Rejects school	40	16	26	40	42	50	47	56	67
Involved with police	17	0	4	6	10	21	19	58	63

*Percentage of all children rated as displaying the characteristic.

Note: These five descriptors are merely a representative sample of the 51 actuarial correlates of the Delinquency Scale.

Source: Material from *Actuarial Assessment of Child and Adolescent Personality: An Interpretive Guide for the Personality Inventory for Children Profile* copyright © 1979 by Western Psychological Services. Reprinted by permission of the publisher, Western Psychological Services, 12031 Wilshire Boulevard, Los Angeles, CA 90025, United States of America.

	Jimmy (DLQ = 114)	Johnny (DLQ = 54)
Refuses to go to bed	42%	18%
Lies	90%	44%
Uses drugs	32%	0%
Rejects school	56%	16%
Involved with police	58%	0%

The reader will immediately recognize that Jimmy fits a pattern of pervasive conduct disorder, whereas Johnny appears to have few such behavior problems. In Jimmy's case, the underachievement is most likely secondary to a pattern of antisocial behavior, whereas for Johnny the clinician must look elsewhere to understand the school failure. Of course, this is only a small fraction of the information that would be available from a computer-based actuarial interpretation of the PIC. In a full report, the clinician would receive statistics and narrative statements pertinent to all 20 scales from the PIC.

Unfortunately, great effort and expense are needed to develop actuarial tables like those provided for the original PIC. Few test publishers are willing to take on the financial burden. Increasingly, test developers rely on clinical judgment as the basis for computer-assisted assessment.

Computer-Assisted Clinical Reports

In a computer-assisted clinical report, the interpretive statements assigned to test results are based upon the judgment of one or more expert clinicians. The expert clinicians formalize their thought processes and develop automated decision rules that are then translated into computer code. This method differs crucially from the computer-assisted actuarial approach in which interpretive statements are based strictly upon formal research findings. Superficially, the two approaches may appear to be identical insofar as each is rule based and automated. The difference has to do with the origin of the rules: empirical research (actuarial approach) versus clinician judgment (clinical approach).

Even though clinicians generally recognize the superiority of the actuarial method, there is one significant advantage to the computer-assisted clinical approach. The advantage is that the clinical approach can be designed to interpret all test profiles, whereas some test profiles will be uninterpretable by means of an actuarial approach. The discouraging truth about actuarial "cookbook" systems for test interpretation is that the classification rate usually plummets when a system is used in a new setting. The classification rate refers to the percentage of test results that fit the complex profile classification rules necessary for actuarial interpretation. For example, in the Gilberstadt and Duker (**1965 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib589)**) actuarial MMPI system, the 1-2-3 code type is defined by these rules for the Hs (Hypochondriasis), D (Depression), Hysteria (Hs), and L, F, K (validity) scales:

- Hs, D, and Hy over *T* score 70

- Hs > D > Hy
- No other scales over *T* score 70
- L < *T* score 66, F < *T* score 86, and K < *T* score 71

Persons who produce this kind of MMPI profile often suffer from psychophysiological overreactivity, not to mention a host of other empirically confirmed characteristics. Of course, there are several additional code types, each defined by a set of complex decision rules, and each accompanied by an elaborate, actuarially based description of personality and psychopathology. A typical finding is that a computer-assisted actuarial system developed within one client population will be capable of interpreting up to 85 percent of the test profiles encountered in that setting. However, when the actuarial system is applied to a new client population, perhaps 50 percent of the test profiles will fit the decision rules. This means that about half of the test profiles do not fit the rules. At best, these clients will receive a superficial, scale-by-scale interpretation rather than a more sophisticated actuarial interpretation based upon code types. The problem of shrinkage in classification rate is observed in virtually all studies of actuarial interpretation **(More-land, 1992 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1173))**.

Computer-assisted clinical reports tend to be lengthy and detailed, full of scale scores, item indices, and graphs. Of course, these reports also include several pages of narrative report, usually phrased in terms of hypotheses as opposed to confirmed findings. The shortest such report is about six pages (e.g., the Karson Clinical Report for the 16PF), whereas longer ones can run to 10 or 20 pages (e.g., MMPI-2 interpretations).

[3]Respectively, the full names for these scales are L (validity scale), Paranoia, Schizophrenia, Hysteria, and Psychasthenia.

12.8 INTERACTIVE VIDEO, VIRTUAL REALITY, AND SMARTPHONES

The digital revolution continues to accelerate, with far reaching consequences for every aspect of society, including psychological assessment. Digital devices such as smartphones, tablets, and laptops have become smaller, faster, cheaper, and possess bandwidth hardly imagined a few years ago. The revolution promises to enhance psychological assessment in ways we are just beginning to comprehend, but also guarantees new ethical challenges. Consider one seemingly inconsequential quandary posed by the digital revolution: Before conducting an assessment, should a psychologist Google a client? Unless personal safety could be at stake, the answer is "No":

> Curiosity about a client is not a clinically appropriate reason to do an Internet search. Let's put it this way: If you know that your client plays in a soccer league, it would be a little odd if on Saturday afternoon you drove by the game to see how your client is doing. In the same way, if you're doing a search, thinking, "What can I find out about this person?" that raises questions about the psychologist's motives **(Martin, 2010 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1053))**.

The promise of the digital revolution is huge, but alongside new developments, psychology and other health professions will need new ethical guidelines.

Interactive Video in Assessment

One recent approach to assessment made possible by modern computer technology is the flexible, interactive presentation of high quality, captivating video segments. At IBM, researchers have been developing the Workplace Situations test to assess job applicants for manufacturing positions **(Drasgow, Olson-Buchanan, & Moberg, 1999 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib435))**. What is unique about the test is the nature of the stimuli. Rather than merely describing work situations, the test displays computer-driven interactive video of realistic work scenes. The assessment consists of 30 short scenes in a fictional organization named Quintronics. The scenes depict work-related interpersonal episodes arising in the manufacture of hypothetical electronic products called quintelles and alpha pin-hole boards. The computer vignettes depict such concerns as excessive workloads, poor training, interpersonal conflict, poor productivity, and flawed work. Each scene is presented and then the screen pauses with a description of five ways of responding to the workplace problem. The scenes have a highly realistic feeling to them, which enhances the face validity of the test. This kind of interactive video test likely provides a more accurate assessment than paper-and-pencil tests of how people would actually respond on the job. Tests that use interactive video are especially good at tapping examinees' abilities to deal with complex, real-life problems, such as decision making under time pressure or conflict resolution in the workplace.

Olson-Buchanan et al. **(1998 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1255))** have developed an interactive video test of conflict resolution that reveals both the promise and the perils of this new technology. Their instrument, the Conflict Resolution Skills Assessment (CRSA), consists of nine conflict scenes, each with the potential for multiple branchings, depending upon the examinee's ongoing response pattern:

> A typical item on the Conflict Resolution Skills Assessment begins by presenting a conflict scene (1–3 minutes in duration) to an individual. At a critical point the scene is stopped and four options for addressing the conflict are provided; the assessee is asked to choose the option that best describes what he or she would do in this situation. Depending on the option chosen, the computer branches to an extension of the first scene depicting how events might unfold. Again, the conflict escalates, the scene is frozen, four options for addressing the conflict are presented, and the assessee decides which option would best resolve the conflict. The computer then branches to an entirely new conflict scene. (p. 180)

The perils of this effort include the increased expense required for test development (e.g., cost of producing high-quality, convincing videos) as well as daunting theoretical issues (e.g., challenge of conceptualizing "good" conflict resolution skills). This kind of interactive, branching, video-based test also poses unique psychometric problems. For example, how do you assess the reliability of specific subelements of the test when only a few of the examinees may have taken that "route" through the test?

In spite of these challenges, the development of path breaking instruments such as the CRSA is well worth the effort. Consider one important payoff, namely, scores on the CRSA show essentially no correlation with general cognitive ability (**Drasgow et al., 1999** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib435)**). Psychologists long have suspected that social skills are distinct from cognitive skills, but when both are assessed with traditional paper-and-pencil instruments, moderate to strong correlations are the rule. Most likely, this is because of shared method variance, namely, verbal test-taking skills help an examinee navigate any paper-and-pencil test, regardless of the construct being measured. By using interactive video as the primary test stimulus, instruments such as the CRSA provide a purer measure of social skills than paper-and-pencil tests. This unique instrument illustrates that social skills contribute something different than cognitive skills to effective work performance.

Another potential application of interactive video is in personnel screening for entry-level police officers. Law enforcement personnel must have good observational and evaluative skills, which can be assessed realistically with video stimuli. For example, an assessment might consist, in part, of a videotape of witnesses at a crime scene. Police candidates might be asked to determine the truth of the witnesses and to draw conclusions about the crime based upon their observational powers (*APA Monitor*, June 1994). This example—currently hypothetical—illustrates the potential for multimedia to revolutionize psychological assessment.

It is worth noting that interactive video tests can be virtually free of reading and writing requirements on the part of the examinee. Talented job candidates who do not possess good reading or writing abilities but who do have practical job skills can be identified by means of these tests. For some jobs, interactive video might be fairer than the paper-and-pencil approach.

Virtual Reality Approaches to Assessment

Virtual reality (VR) is a sophisticated mode of human-computer interface that allows users to navigate and manipulate three-dimensional environments in a naturalistic fashion (**Vince, 2004** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1700)**). The participant wears a pair of goggles that transmit realistic, three-dimensional images of a simulated environment, called a virtual environment (VE). In the most sophisticated applications of VR, the user also would wear gloves that are interfaced with the video display so that objects in the VE can be manipulated. More commonly, especially in psychological assessments with VR, the simulated environment is navigated with a joystick or similar device. The VR user can walk, run, or even fly through the VE and explore points of reference that would be difficult or impossible in the real world (**Vince, 2004** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1700)**).

New assessment tools that utilize virtual reality are in their infancy, but show great promise. One positive feature of these tests is that most possess good ecological validity. The required tasks highly resemble real world issues and concerns. Consider the contrast between a paper-and-pencil measure and a VR measure of executive functions. The Trail-Making Test, part B (TMT-B), requires the user to draw a line between numbers and letters in alternating order as quickly as possible (**Reitan & Wolfson, 1993** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1350)**). It is considered a measure of executive functions, among other skills. The VE Grocery Store test requires the user to navigate a VR-simulated grocery store in search of shopping list items (**Parsons, Rizzo, Brennan, Bittman, & Zelinski, 2008** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1268)**). This, too, is considered a measure of executive functions. The TMT-B may be a good test, but the task demands are foreign to everyday living. In real life, we *never* connect numbers and letters with a pencil line. In contrast, almost everyone has the need to shop for groceries. The VE Grocery Store test embodies good ecological validity. The tasks of the test include:

- navigating through a virtual grocery store by following specified routes though the aisles
- finding and selecting items needed to prepare simple meals, such as making a peanut butter and jelly sandwich
- pricing and selecting other items so that no more than a budgeted amount is spent
- performing a prospective memory task when a certain individual is encountered.

Researchers can vary the density of items on the shelves, the similarity of packaging, and the background distractions (e.g., loudspeaker announcements). Because of its strong ecological validity, the VE Grocery Store test produces "subjective engagement that is equivalent to engagement in the real world (**Parsons et al., 2008** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1268)** , p. 187)." While promising, the test is currently an experimental measure in need of further validation.

Buxbaum, Dawson, and Linsley (**2012** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib249)**) recently validated a virtual reality measure of hemispatial neglect, suitable for patients who have sustained a right-hemisphere stroke. In 40 to 50 percent of cases, right-hemisphere stroke patients demonstrate impairment in the detection of objects, persons, and events on the left side of space, a condition known as hemispatial neglect. This condition constitutes a serious barrier to independence. For example, patients tend to bump into things when walking. While paper-and-pencil tasks such as line bisection are useful in diagnosis, they do not detect subtle cases of the disorder.

Dawson, Buxbaum, and Rizzo (**2008** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib401)**) developed the Virtual Reality Lateralized Attention Test (VRLAT) to provide an ecologically valid and more sensitive measure of hemispatial neglect (**Dawson, Buxbaum, & Rizzo, 2008** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib401)**). The test requires patients to travel along a virtual path using a joystick (participant condition) or passively viewing the environment while the examiner controls the pace (examiner condition). The task of the patient is to identify virtual objects on both sides of the path and to avoid collisions with the objects. As patients negotiate the path, they are instructed to call out the objects they see, including unique attributes (e.g., blue tree, pig statue, camel statue, orange tree). Each path is negotiated in both directions, so every object occurs once on the left side, and once on the right side. In total, users complete six pathways. Individual responses are scored 0 to 3 depending on whether the object is detected and correctly described. Of course, there are separate scores for left-sided objects and right-sided objects. A disparity between the scores favoring the detection of right-sided objects would be interpreted as left-sided hemispatial neglect. This is precisely what the researchers found in a sample of 64 post-acute right-hemisphere stroke patients. Based on appropriate statistical analyses, the researchers concluded that the VRLAT possesses strong sensitivity and specificity, minimal practice effects, and strong validity (**Buxbaum et al., 2012** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib249)**). Further, it was a better predictor of real-world collisions than a battery of paper-and-pencil tests. The collision test involved patients quickly traversing a 150 foot long maze with multiple left and right turns. The maze corridor was 3.5 feet in width. The number of collisions with the walls was recorded.

For purposes of illustration, we have described here only two assessment tools that utilize virtual reality, The VE Grocery Store test and the Virtual Reality Lateralized Attention Test. Dozens more VR tests are available or in production. Innovative approaches based on VR can be found in several journals, including *Virtual Reality* and *CyberPsychology, Behavior, and Social Networking*.

12.9 EVALUATION OF COMPUTER-BASED TEST INTERPRETATION

Computerized testing has clear advantages but also some potentially serious disadvantages in comparison to the traditional clinical approach to psychological testing. We offer a brief survey here, stressing both the advantages and disadvantages of computer-based testing, diagnosis, and report writing. More detail on this topic can be found in Butcher (**1987** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib242)**), Moreland (**1992** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1173)**), Roid and Johnson (**1998** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1401)**), Butcher, Perry, and Atlis (**2000** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib243)**), and Mills, Potenza, Fremer, and Ward (**2002** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1162)**).

Advantages of Computerized Testing and Report Writing

The main advantages of computer-based testing are quick turnaround, inexpensive cost, near-perfect reliability, and complete objectivity. In addition, some measurement applications such as flexible adaptive testing virtually require the use of computers for their implementation. We explore these points in more detail later.

In a busy clinical practice, delays between testing and submission of the consulting report are common, almost inevitable. These delays not only tarnish the reputation of the consultant, but they may also adversely affect the treatment outcome for the client. For example, a college student with learning disabilities may need immediate intervention in order to avert an academic disaster. A delay of two or three weeks in submission of a consulting report could spell, indirectly, the difference between failure and success in academic performance. Computer-based reports can speed up the entire consultation process. Many software systems produce reports that can be transferred into a standard word-processing program for immediate customized editing, thereby speeding up the turnaround time (e.g **Psychological Corporation, 1994** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1325)** ; **Tanner, 1992** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1611)**).

Cost is another consideration in computer-based testing. Although there are no definitive studies on this topic, most authorities assert that computer-scored and interpreted psychological tests cost considerably less than those produced entirely by clinician effort (**Butcher, 1987** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib242)**). In their studies of automated testing at the Salt Lake City VA Hospital, Klingler, Miller, Johnson, and Williams (**1977** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib903)**) concluded that the computer cut the cost of testing in half. Certainly as the computerized testing programs become more sophisticated and are used by larger numbers of clinicians, the cost per consultation will plummet.

Reliability and objectivity are the hallmarks of the computer. Assuming that the software is accurate and error-free, computers simply do not make clerical scoring errors, nor do they vary their methods of stimulus presentation from one day to the next, nor do they yield different narrative reports based on the same input. The product is the same no matter how many times the computer program is used. Furthermore, because computerized reports are based on objective rules, they are not distorted by halo effects or other subjective biases that might enter into a clinically derived report. Butcher (**1987** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib242)**) asserts that computerized reports could have special significance in court cases, because they would be viewed as "untouched by human hands." This is an intriguing possibility, but perhaps somewhat overly optimistic. Lawyers and judges will still want to know who programmed the software, how the narrative statements were developed, and so on.

Disadvantages of Computerized Testing and Report Writing

Consider the following illustration, hypothetical yet realistic and probably not a rare occurrence. A hospital physician refers a difficult medical patient to the psychology service for a personality evaluation. The patient is escorted to the testing center where a receptionist seats him at a table in front of a microcomputer. Instructions

appear on the computer monitor to answer a series of self-statements true or false by pressing the T or F key. The patient completes the computerized objective personality inventory and is escorted back to the medical service. Seconds later, a narrative report based on the patient's responses emerges from the printer. The consulting psychologist peruses the report briefly, then sends it (unsigned) through departmental mail to the physician. The report is handsome, ever so crisp in its laser-printed appearance, with a graphic summary of scales on the cover page. Furthermore, the narrative is valid sounding and reads as if it were copyedited by a professional writer (in fact, it was). The physician is impressed and takes the report to heart, making treatment decisions based on the personality evaluation.

This scenario illustrates an essential quandary with computer-based testing and report writing: Computers can so dominate the testing process that the clinical psychologist is demoted to a mere clerk—or is removed from the assessment loop entirely. Although most psychologists acknowledge that computers are a welcome addition to the practice of psychological testing, critics have raised a number of disquieting concerns about recent assessment practices such as those depicted here. Computerization of the testing process raises practical, legal, ethical, and measurement issues that deserve thoughtful review.

In general, skeptics do not attack the practice of computerizing the mechanics of test administration and scoring; these computer applications are seen as efficient and appropriate uses of modern technology. Nonetheless, even the most ardent proponents acknowledge the need to investigate test-form equivalency when an existing test is adapted to computerized administration. In particular, practitioners should not assume that the computerized adaptation and the original version of a test produce identical results. Equivalency is an empirical issue that must be demonstrated by appropriate research. For most tests, equivalency can be demonstrated, but this must not be taken for granted (**Lukin, Dowd, Plake, & Kraft, 1985**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1014) ; **Schuldberg, 1988**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1475)).

Some tests do not maintain score equivalency when translated to computer. The Category Test (CT) from the Halstead-Reitan Neuropsychological Battery is a case in point. In a comparison of computerized and standard versions of the Category Test with rehabilitation patients, Berger, Chibnall, and Gfeller (**1994**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib147)) found a huge difference in error rate for two groups of subjects who had equivalent backgrounds: an average of 84 errors on the computerized CT versus an average of 66 errors on the standard CT test. Apparently, the computerized CT test is much more difficult than the standard version, which means that separate norms must be developed for its interpretation. Much smaller differences between computerized and standard test administration have also been reported for the MMPI, with computer-based scores tending to underestimate (very slightly) the booklet-based scores (**Watson, Thomas, & Anderson, 1992**
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1723)).

12.10 COMPUTERIZED ADAPTIVE TESTING

A final advantage of computer-based testing is its application to flexible adaptive testing. Adaptive testing is nothing new—Binet used it when he worked out the methods for finding the basal and ceiling items on his famous intelligence test. Binet placed his items along a continuum of difficulty so that the examiner could test downward to find the examinee's basal level and test upward to find the ceiling level. This procedure eliminated the need to administer irrelevant items—those so easy (below the basal level) that the examinee would surely pass them, or those so hard (above the ceiling level) that the examinee would surely fail them. Another example of adaptive testing is the two-stage procedure whereby results on an initial routing test are used to determine the entry level for subsequent scales. For example, on the Stanford-Binet: Fifth Edition, results of the initial vocabulary and matrices subtests determine the starting points for subsequent subtests. By reducing the time needed to obtain an accurate measure of ability, adaptive testing fulfills a very constructive purpose.

Computerized adaptive testing (CAT)
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss59) is a family of procedures that allows for accurate and efficient measurement of ability (Wainer, 2002). Although details differ from one method to another, most forms of computerized adaptive testing share the following features:

- Based on extensive pretesting, the item response characteristics of each item (e.g., percentage passing versus ability) are appraised precisely.
- These item response characteristics and a CAT item-selection strategy are programmed into the computer.
- In selecting the next item for presentation, the computer uses the examinee's total history of responses up to that point.
- The computer recalculates the examinee's estimated ability level after each response.
- The computer also estimates the precision of measurement (e.g., standard error of measurement) after each response.
- Testing continues until a predetermined level of measurement precision is reached.
- The examinee's score is based on the difficulty level and other measurement characteristics of items passed, not on the total number of items correct.

The measurement advantages of CAT can be summarized in two words: precision and efficiency **(Weiss & Vale, 1987 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1745)**). Regarding precision, CAT guarantees that each examinee is measured with the same degree of precision because testing continues until this criterion is met. This is not so with traditional tests in which scores at both tails of the distribution reflect greater levels of measurement error than scores in the middle of the distribution. Regarding efficiency, the CAT approach requires far fewer test items than are needed in traditional testing. For example, written certification examinations usually include 200 to 500 items, while CAT examinations are always shorter, often including fewer than 100 items to achieve a more accurate level of measurement **(Lunz & Bergstrom, 1994 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1015)**). In one analysis, the reliability of alternative computer-adaptive tests for certification in medical technology was .96 **(Lunz, Bergstrom, & Wright, 1994 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1015)**). This is remarkable because shorter tests (the goal in CAT testing) tend to have lower reliability than longer tests (such as found in traditional testing programs).

In addition to increased measurement efficiency, CAT has many other advantages over traditional paper-and-pencil assessment. Wainer **(2000 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1704)**) points out that CAT allows for better test security, immediate scoring and feedback, equal challenge to all examinees, presenting of new items, and the use of a variety of question types.

Regarding the last point, examples of novel item types not possible on a traditional multiple-choice exam include spoken words (such as for a spelling test), open-ended math problems (the answer is typed in), and video segments (followed by written questions).

The CAT approach to psychological testing has been used mainly by large organizations such as the U.S. Army and the Educational Testing Service for assessment of intelligence and special abilities. In recent years, national licensing boards (e.g., in medicine) have begun to implement CAT testing because of convenience in scheduling tests, tighter control over test security, reduced costs, and the opportunity for better data collection (**Lunz & Bergstrom, 1994** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1015)). Technical information on CAT systems is proprietary and difficult to obtain. Nonetheless, it is clear that the efficiency of the CAT approach is substantial. CAT uses fewer items of better quality than a conventional test of the same length. A general finding is that CAT reduces test length by about 50 percent, with reductions for individual examinees of up to 80 percent, with no loss in measurement accuracy (**Laatsch & Choca, 1994** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib937) ; **Weiss & Vale, 1987** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1745)).

One recent study revealed phenomenal success for the CAT approach in reducing the time spent in testing, and simultaneously providing better discriminant validity in the assessment of depressive symptomatology (**Gibbons, Weiss, Kupfer, Frank, Fagiolini, and others, 2008** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib586)). The study involved 800 outpatients who completed the 616-item Mood and Anxiety Spectrum Scales (MASS) at two times. The first administration was used to develop and evaluate a CAT version of the MASS, while the second administration confirmed the functioning of the CAT method in live testing. The CAT version utilized an average of *95 percent* fewer test items (30 instead of 616) and yet was shown to provide *better* discrimination of seriously depressed versus mildly depressed patients.

Recently, CAT has been applied to mainstream personality tests like the MMPI-2 with encouraging results. Forbey and Ben-Porath (**2007** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib523)) provide a review of the MMPI-2 computerized adaptive version. They conclude that the new approach provides the same accuracy of measurement with about a 20 percent reduction in the number of items administered. Even so, CAT approaches with the MMPI-2 are experimental at this point, and not likely to receive significant clinical usage for several years.

There may well be reasons for caution in the application of CAT to personality testing. An unavoidable consequence of using CAT is that item order will change from one examinee to another, which may invoke context effects that influence subsequent item responding. Ortner (**2008** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1257)) investigated this prospect with a CAT version of the Eysenck Personality Profiler (EPP) with 362 German adults. Some participants first encountered items representing extreme trait levels, while others were first exposed to items representing medium trait levels. These initial exposures distorted their subsequent responses to the point where scores on three out of seven scales of the EPP shifted up or down significantly. These findings indicate that context effects may be a problem in using CAT with personality inventories.

As the cost of computing continues to plummet, more and more large-scale applications of CAT will be developed. In the late 1990s, the Educational Testing Service moved toward near total reliance on CAT versions of the Graduate Record Examination and other selection tests. Licensing and certification boards such as the National Council of State Boards of Nursing also have introduced CAT versions of their certification tests. Mills and Stocking (1996) discuss practical issues in large-scale computerized testing.

12.11 THE FUTURE OF TESTING

What is the future of psychological testing in the twenty-first century? We will hazard a few speculations here, cognizant that prognostications about the future often are wrong. Forecasting developments in testing is especially difficult because the enterprise is increasingly constrained, directly or indirectly, by public opinion. For example, at one point in the 1980s the legislature of the state of California made it *illegal* for school psychologists to use traditional intelligence tests as a basis for placing minority students in special education classes. These restraints on testing were driven by public outrage over the excessive placement of minority students in special education classes. Thus, even when a particular technology of testing is feasible and promoted by psychologists, there exists the possibility that it might be strictly controlled or even banned.

A case in point is Matarazzo's (**1992** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1060)**) prediction that biological measures of intelligence will gain prominence in the twenty-first century. Certainly it appears true that biological measures of ability such as averaged evoked potential (gauged from EEG waves), or glucose metabolic rate in the brain (gauged from PET scans), or relative brain size (gauged from MRI scans) will prove to be effective approaches to assessment (see **Topic 5A** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch05#ch05box1)** , Theories of Intelligence and Factor Analysis). But Matarazzo (**1992** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib244)**) goes further in asserting that these and other biological approaches actually will receive common usage:

> Therefore, another of my predictions is that in the early decades of the 21st century we may see the further development and *use in practice* of these and other biological indices of brain function and structure in a test (or a test battery) for the measurement of individual differences in mental ability, thus heralding the first clear break from test items and tests in the Binet tradition in a century. (p. 1012, italics in the original)

Although Matarazzo's prediction could come true, a more likely scenario is that the general public will be threatened when biological indices are used in assessment and will, therefore, take steps (e.g., pressure on legislators) to ensure that such measures receive limited (if any) application. The public will be threatened because, rightly or wrongly, biological characteristics such as glucose metabolic rate in the brain are perceived to be relatively permanent and immutable. The fear will arise that biological tests will sort people into a caste system. Even if (or when) the validity of biological tests is firmly established, it will be decades (if ever) until they are found acceptable by the general public.

Trends in Testing: A Few Confident Predictions

The computerization of testing is already a fixture of industrialized societies and this trend can only increase in the future. Existing tests will be adapted to the desktop computer with increasing regularity. An example of this trend is Fepsy (Ferrum 1 Psyche), a system for automated neuropsychological testing that is available online at 220 sites throughout the Netherlands and most of Europe. Fepsy is described on the Internet at **www.euronet.nl/users/fepsy** **(http://www.euronet.nl/users/fepsy)** . Fepsy consists of the following subtests:

- Auditory reaction time
- Binary choice reaction time
- Tapping task
- Visual searching task
- Recognition tasks
- Vigilance task
- Rhythm task
- Classification task
- Six Visual Half field tasks
- Corsi block tapping

A common use is pre- and postoperative testing of patients who undergo epilepsy surgery for relief of seizures. The system has even been used with fully conscious patients *during* surgery. Under local anesthesia, the patient works on a subtest while simultaneously receiving harmless electrical stimulation at distinctive sites on the cortex. The purpose is to determine whether specific cognitive functions might be affected when scar tissue is excised from the brain. The advantage of using a multicenter, multinational, computerized testing system is that the examiner has access to normative data for *thousands* of patients with specific conditions.

Another prediction is that fewer and fewer wide-spectrum tests (e.g., personality inventories and individual intelligence tests) will be released by test publishers (**Gregory, 1998** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib649)**). Instead, publishers will concentrate on tests designed to assess particular areas of functioning for specific target populations (e.g., measures of memory functioning for elderly persons suspected of having dementia). The reasons for these complementary trends are economic:

> Test publishing is big business, a respectable way for large corporations to earn a profit. Publishers will be reluctant to make the major investment needed to develop new instruments that have the grandiose ambition of assessing many aspects of personality or intellect for a wide range of subjects. The cost is too high and—in light of the existing competition—the risk is too great. (**Gregory, 1998** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib649)** , pp. 76–77)

Test publishers likely will focus on less-expensive and less-risky forms of test development such as instruments that embody distinctive constructs relevant to specific target groups. Examples might include tests to measure risky behaviors in adolescents, mental decline in elderly persons, faulty cognitions in depressed persons, or communication problems in maritally distressed couples. These kinds of instruments will flourish, whereas publishers will rarely invest in new omnibus tests of personality or ability, preferring instead to revise and recycle existing instruments.

We can also predict with some confidence that the movement toward evidence-based assessment will gain strength in the years ahead. In **evidence-based assessment** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss101)** , the soundness of a testing tool is evaluated not just by means of the standard psychometric indices of reliability and validity but through considerations of clinical utility as well (**Barlow, 2005** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib91)**).

Clinical utility is a broad and imprecise concept that consists of several features, including treatment utility, monetary cost, psychological cost, and client acceptability (**Hunsley & Mash, 2005** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib799)**). Treatment utility is vital. This is the extent to which assessment data contributes to positive treatment outcomes. Does the client get better, faster, as a result of the assessment? If not, what is the point? But better outcomes are not the only consideration in clinical utility. The monetary cost of assessment needs to be considered as well. Even a helpful assessment may prove to be counterproductive if the cost is prohibitive. Someone has to pay for assessments, whether it is the clients or the insurance companies. More money for assessment means less money is available for other purposes such as ongoing psychotherapy. Cost is always an issue in health care (**Cummings, 2007** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib379)**). Clinical utility also includes the psychological cost of measurement errors. For example, when assessment *incorrectly* suggests that an older adult shows signs of dementia, the client and family pay an emotional price. Likewise, false negative results (e.g., concluding that a patient with mild dementia is normal) also exact an emotional toll. Emotional costs are intangible, but nonetheless an important consideration in clinical utility. Finally, client acceptability needs to be considered as well (**Hunsley & Mash, 2005** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib799)**). Will the client agree to complete the assessment? Or will the client resist the assessment and thereby produce misleading results? All of these factors need to be considered in clinical utility.

Driven in large measure by the insistence of the insurance industry that therapies must be empirically based, there is a growing demand for brief, effective treatments throughout the entire field of health care. Evidence-based

assessment is inescapably intertwined with this national movement toward evidence-based therapy for medical and psychological illnesses. Side by side with this trend, we can expect to see a keen emphasis upon empirically based psychological assessments.

Finally, we can predict that positive psychological assessment will gain greater popularity. **Positive psychological assessment** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss248) is a natural spin-off of the positive psychology movement, which is defined as "the scientific and practical pursuit of optimal human functioning" (**Lopez & Snyder, 2003** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1004)). Proponents of the positive psychology movement find the current focus of assessment—with its emphasis on pathology and what is wrong with people—to be lopsided and incomplete. A full understanding of persons also includes an appraisal of with is *right* with them. It includes a census of such positive qualities as hope, creativity, wisdom, courage, forgiveness, humor, gratitude, and coping. The traditional instruments of psychological testing—for example, the Rorschach, MMPI-2, MCMI-III, and so forth—provide essentially no information on these positive human qualities. In the years ahead, new instruments and original philosophies of assessment will most certainly redress the imbalance.

The Smartphone Revolution

By 2025, more than 5 billion people will be using ultra-broadband smartphones with capacities far beyond current technology (**Miller, 2012** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1147)). Although smartphones were not designed for purposes of psychological testing, they possess the potential for implementing a large range of ecologically valid assessments. To envision the smartphones of the future, we need a phenomenal leap of imagination. We also need new vocabulary, specifically, we need to know the definition of teraflop. This is a measure of computer speed and refers to a trillion decimal number computations per second, in other words, *really fast*. In 2025, smartphones are likely to possess

> at least eight 200 GHZ [gigahertz] processors, yielding about 10 teraflops—making them ten times faster than the first teraflop supercomputer in 1997, the $50 million, 9,600 processor Intel ASCI Red that filled a whole room. Such powerful smartphones cold run complex psych apps continually in the background (e.g., running emotion detection algorithms on voice input or combining GPS and geographic information system (GIS) data into measures of daily movement patterns), without disrupting other apps and annoying participants. Future smartphones—basically handheld supercomputers—will be able to run psych apps of nearly limitless complexity (**Miller, 2012** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1147) , p. 224).

Future applications will extend well beyond simple surveys and tests. Once clients download the appropriate "testing apps," smartphones could be programmed for countless forms of creative assessment. Here are a few examples of current and future assessment uses (**Miller, 2012** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1147)). In each case, the permission of the user would be needed:

- estimating fast food exposure from a user's GPS and GIS data
- gathering data from biosensors for remote physiology assessment
- analyzing social context and behavior from ambient sounds
- correlating mood reports with GPS locations and ambient noise levels
- analyzing voice modulation for patterns of stress

With the enlistment of the users, a wide range of ecological momentary assessments could be gathered (**Courvoisier, Eid, & Lischetzke, 2012** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib363)). One example: Migraine headache sufferers could provide pain ratings at random intervals throughout the day, to determine the efficacy of treatments. The smart-phone might chime a distinctive tone, signaling a request to tap an on-screen scale. Completing the scale would take about five seconds. Another example: Short surveys could be administered following phone calls from significant others to identify patterns of emotional reactivity.

While holding great promise, smartphones in assessment also carry likely pitfalls (**Miller, 2012** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1147)**). Obtaining truly informed consent will be more problematic because users do not read software licensing agreements before clicking "I agree." Further, confidentiality is difficult to guarantee because of the vulnerability of digital systems to hacking. Liability is another concern. Developers of testing apps could be liable for unintended consequences. A programming bug might cause smartphones to malfunction or could prevent emergency calls. Further, technology is changing so fast that established practitioners will need constant updating:

> How can older researchers grow comfortable with such a futuristic technology—one that is, to all intents and purposes, indistinguishable from magic? (**Miller, 2012** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1147)** , p. 234).

Smartphone applications in assessment will raise challenging ethical and practical issues. Even so, the future is bright in this new arena of testing.

Testing and the Next Big Questions in Psychology

In this closing section of the book, we turn to some admittedly more speculative predictions about the future of testing. In doing so, the reader is invited to exercise his or her imagination as well. After all, psychological testing is here to stay, and will continue to evolve and adapt. As it has for more than a century, testing will continue to play a significant role in psychology and modern society. But exactly how?

The starting point for this final conversation is a fascinating issue of *Perspectives on Psychological Science*, a journal of the Association for Psychological Science (**Diener, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib415)**). The journal editor asked a number of leading psychologists—none from the field of psychometrics—to write about the most important questions to be asked in their particular specialty in the upcoming decade. These questions have been reproduced in paraphrased manner (for consistency and clarity) in **Table 12.8** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec11#ch12tab8)** . A few esoteric contributions have been omitted.

Of course, the whole point in asking a question is to hope that an answer will be found. While at first glance it may seem only a slight possibility that psychological testing could contribute to answering any of these questions, on closer examination it seems likely that testing will play an essential role in many cases.

Consider the question of nature and nurture of plasticity in early human development (**Belsky & Pluess, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib130)**). This is a general topic that invites many specific lines of inquiry. For example, Davis et al. (**2007** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib395)**) found that prenatal maternal depression and elevated cortisol levels in late pregnancy predicted negative reactivity in children at age 2. But what is "negative reactivity?" The dependent variable in this line of research—reactivity in children—is a construct measured by rating scales and situational tests. Thus, one line of answers to the underlying question ("What is the nature and nurture of plasticity in early human development?") likely will hinge on the development of precise and valid measures of reactivity in children. This is a clear role for psychological testing in answering one of the big questions in psychology.

TABLE 12.8 The Next Big Questions in Diverse Fields of Psychology

What is the connection between complex psychological states such as emotion or cognition and the physical substrates of the brain? (**Barrett, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib96)**)

Why do people do what they do? And, what are the important situational and personality variables in answering this question? (**Funder, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib553)**)

What is the nature and nurture of plasticity in early human development? (**Belsky & Pluess, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib130)**)

How do we achieve a synthesized understanding of early cognitive development from studies of separate cognitive abilities? (**Oakes, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1246)**)

How can evolutionary psychology successfully explain personality and individual differences? (**Buss, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib237)**)

How do stressful events and negative emotions influence the immune system, and how big are the effects? (**Kiecolt-Glaser, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib885)**)

How can you tell if a particular memory belonging to you or someone else is true or false? (**Bernstein & Loftus, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib150)**)

Can we improve our physical health by altering our social networks? (**Cohen & Janicki-Deverts, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib319)**)

How can decision making be improved? (**Milkman, Chugh, & Bazerman, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1144)**)

How can we promote self knowledge ("Know thyself") and what are the results of greater self knowledge? (**Wilson, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1768)**)

Can psychological research on correcting cognitive errors promote human welfare? (**Lilienfeld, Ammirati, & Landfield, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib989)**)

Is it possible to teach intuition and can it be enhanced by virtual simulation? (**Seligman & Kahana, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1481)**)

What are the mechanisms of gene-environment interaction effects in the development of conduct disorder? (**Dodge, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib422)**)

Why do different individuals progress along different life trajectories? (**Smith, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1510)**)

How can we live well? How can we achieve and sustain a good life? (**Park & Peterson, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1266)**)

What is the near and distant future of human-android interaction? (**Roese & Amir, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1380)**)

Another question on the list pertains to evolutionary explanations of personality and individual differences (**Buss, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib237)**). Regardless of the particular directions pursued in this line of research, accurate measurement of *appropriate* personality constructs will be required. By *appropriate*, we are referring here not to just any old personality constructs, but to those that "cut nature at its joints." In other words, the personality constructs measured in evolutionary psychology need to capture essential underlying elements of personality that could be susceptible to evolutionary influence. As an example, consider research on humor styles, discussed in an earlier chapter. Using the Humor Styles Questionnaire (**HSQ, Martin, et al., 2003** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1048)**) in a behavior genetics analysis of identical and fraternal twins, Vernon et al. (**2008** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1696)**) found that positive forms of

humor (Affiliative and Self-enhancing) revealed significant genetic influence whereas negative forms of humor (Aggressive and Self-defeating) arose from environmental influences. The point of this digression is that the HSQ successfully partitions humor into meaningful elements, including some that can be explained in evolutionary terms. For example, we could hypothesize that Affiliative humor promotes group bonding which, in turn, promotes individual survival and thus allows for genes to be passed on. This conclusion is possible because of the careful analysis of humor implicit in the development of the relevant personality test, the Humor Situations Questionnaire.

Finally, consider the question whether we can improve our physical health by altering our social networks (**Cohen & Janicki-Deverts, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib319)**). There is no doubt that membership in a diverse social network is correlated with a variety of positive health outcomes, such as resistance to cognitive decline with aging, better prognosis when facing chronic illnesses, and even greater resistance to infectious disease (**Cohen & Janicki-Deverts, 2009** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib319)**). But these results are correlational, not necessarily causal. The pressing question is: if individuals *alter* their social networks will this improve physical health? From the standpoint of psychological testing, the construct of social network is crucial to the answer. What is a social network? How is it assessed or measured? Research in this area of behavioral medicine will require the development of straightforward and valid measures of social network, another role for testing in answering one of the big questions of psychology.

As a final challenge, the reader is invited to review the list in **Table 12.8** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec11#ch12tab8)** . What roles can you see for psychological testing in answering these big questions?