Minnesota Multiphasic Personality Inventory—Adolescent (MMPI-A)
The Minnesota Multiphasic Personality Inventory—Adolescent (MMPI-A) (Butcher et al., 1992) is a 478-item true-false, self-report inventory designed for use with adolescents ages 14–18 years to assess some of the major patterns of personality and emotional disorders. The derived scales are very similar to the MMPI-2 scales listed in Table 8.8. Items measure 6 Validity Scales, 10 Clinical Scales, 15 Content Scales, 6 Supplementary Scales, and about 30 Harris-Lingoes scales. As with any test, it is essential that any statements from computerized sources be validated with other clinical information. The normative sample \((n = 1,620)\) was very diverse, although it may have oversampled a more educated population. It consisted of male \((n = 805)\) and female \((n = 815)\) adolescents ages 14–18 years living in eight U.S. states; one state’s sample was from an American Indian reservation. There was also a large adolescent clinical population \((n = 703)\). Most of these subjects were paid to complete the test (Butcher et al., 1992). This inventory requires a 6th-grade English reading level.

Raw scores are converted to Uniform T percentile-comparable scores for interpretation through use of convenient profile forms. Different scoring keys are used according to gender. The MMPI-A may take up to three hours to complete and can be scored by hand or computer. It is a Level C instrument. Sample items include “I’m afraid to go home,” “Others do not really love me,” and “I feel uneasy outdoors.” Test-retest reliability results range from 0.65 to 0.84 for the Clinical scales (Butcher et al., 1992). Strong internal consistency coefficients were reported for 4 of the 15 basic and clinical scales \((r = 0.80+)\); 7 of 15 were between \(r = 0.60\) and 0.80. Two response set indicators (VRIN and TRIN) are validity scales that show a respondent’s patterns of responding in an inconsistent or contradictory manner (Butcher et al., 1992). The MMPI-A is one of the only adolescent clinical inventories to comprehensively incorporate a number of validity scales to evaluate client response sets (Archer & Krishnamurthy, 2002). Unfortunately, fewer MMPI-A items demonstrate the same discriminative value in differentiating clients from normal and clinical samples than the adult version of the test (Archer & Handel, 2001).

Bright 12- and 13-year-olds can also be tested, as well as 18-year-olds who have completed high school (Lanyon, 1995). As a Level C instrument, examiners are required to undergo training and supervision prior to administration, scoring, and interpretation of this test (Butcher et al., 1992). The MMPI-A has a number of unique features appropriate for its intended use with adolescents, yet several of the scale labels seem outdated and/or offensive (i.e., Masculine-Feminine, Hypomania, Hysteria, and Psychopathic Deviate) (Claiborn, 1995). “Clinicians should recognize that not all adolescents have the necessary skills to complete the MMPI-A” if their reading comprehension skills are inadequate or if their cultural background and life experiences are out of the range of the test (Butcher et al., 1992, p. 27). (Special learning problems and English as a second language may prohibit the prerequisite reading comprehension, including idioms or other cultural meanings.) It may be prudent to break the testing up into smaller sessions because some adolescents may be too easily distracted or unable to complete the test in one sitting (Butcher et al., 1992). The MMPI-A is a good tool that can help to measure psychopathology in adolescents (Archer & Krishnamurthy, 2002; Claiborn, 1995) and is very useful in planning, directing, and evaluating treatment (Lanyon, 1995).

References:


Millon Adolescent Clinical Inventory (MACI)

The Millon Adolescent Clinical Inventory (MACI) (Millon, Millon, & Davis, 1993) is a 160-item inventory that requires a 6th-grade reading level. The MACI is designed to assess an adolescent’s personality, along with self-reported concerns and clinical syndromes using 27 content scales and 4 response bias scales: Personality Patterns, Expressed Concerns, Clinical Syndromes, and Modifying Indices. For further breakdown of the scales, see Table. These scales coordinate with descriptive characteristics in recent DSM classifications (Millon et al., 1993). The test was normed using 13- to 19-year-olds. The development sample (n = 579) was 54% male and 46% female. The two cross-validation samples (n = 139, n = 194) were 53% and 65% male, respectfully, and 47% and 35% female, respectively (Millon et al., 1993). Over 1,000 adolescents and their clinicians from 28 states and Canada were involved in the development of the MACI.

The MACI usually requires about 20 to 40 minutes to complete and can be scored by hand in about 20 minutes, sent to the publisher by mail, or scored by computer onsite in about 5 minutes (Erford, 2006). Sample items include “I have an attractive body,” “I go on eating binges frequently,” and “I enjoy fighting.” Internal consistency reliabilities for the Development Sample range from 0.73 for the Scales D (Sexual Discomfort) and Y (Desirability) to 0.91 for Scale B (Self-Devaluation). Except for Scale VV (Reliability) scores, raw scores are converted to Base Rate Scores (BRS) for interpretation. Different BR transformation tables are used depending on the age and gender of the adolescent and are adjusted to a value that falls between 1 and 115 (Millon et al., 1993). Internal consistencies for the two cross-validation samples combined ranged from 0.69 for Scale D (Sexual Discomfort) to 0.90 for Scale B (Self-Devaluation). Internal consistency coefficients for the development sample Personality Patterns scales ranged from 0.74 for Scale 3 (Submissive) to 0.90 for Scale 8B (Self-Demeaning). Test-retest reliability results ranged from 0.57 for Scale E (Peer Insecurity) to 0.92 for Scale 9 (Borderline Tendency) for a 3- to 7-day interval. The median stability coefficient is reported as 0.82 (Millon et al., 1993). Criterion-related validity correlations are moderate in magnitude (Erford, 2006).

The MACI is designed for use with emotionally disturbed adolescents ages 13–19 years as an aid to help identify, predict, and understand some of the psychological difficulties this group experiences. Since this is a Level C instrument, examiners are required to have “a graduate degree in psychology or a related field, or appropriate licensure, a course in testing theory, coursework in personality theory, or abnormal psychology, and appropriate experience under supervision” (Erford, 2006, p. 41). Strengths of the MACI include ease of scoring and interpretation, personality variables mapped to DSM personality disorders, appropriateness of concerns frequently expressed by emotionally disturbed adolescents, and identification of important clinical syndromes (Retzlaff, 1995). Clinicians using the computer interpretive report are likely to find the response cover sheet, printout, histogramic display, narrative, and list of correlated Axis I and II entities useful (Stuart, 1995). Weaknesses of the MACI include the underrepresentation of participants ages 18–19 years in the normative samples (Stuart, 1995). The manual clearly stated that use of the MACI for any population outside the 13–19 age designation would be inappropriate (Millon et al., 1993). There is a lack of item and scale specificity because 160 items attempt to score 30 scales (Retzlaff). Also, overrepresentation of Whites (78.8%) (Stuart) and males in the normative sample may make it less appropriate for use with some populations (Millon et al., 1993). Lastly, it may not be particularly useful as a screening level test for the general adolescent population because the norming sample did not include adolescents not identified as patients in treatment programs (Stuart, 1995). Overall, the best use of the MACI is for hypothesis generation and validation, outcomes assessment, and screening for pathology, not for diagnosis.
Table. MACI Response bias scales and content scales

<table>
<thead>
<tr>
<th>Personality patterns</th>
<th>Expressed concerns</th>
<th>Clinical syndromes</th>
<th>Modifying indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale 1—Introversive Scale</td>
<td>A—Identity diffusion</td>
<td>Scale AA—Eating dysfunctions</td>
<td>Scale X—Disclosure</td>
</tr>
<tr>
<td>Scale 2A—Inhibited</td>
<td>Scale B—Self-devaluation</td>
<td>Scale BB—Substance-abuse proneness</td>
<td>Scale Y—Desirability</td>
</tr>
<tr>
<td>Scale 2B—Doleful Scale</td>
<td>C—Body disapproval</td>
<td>CC—Delinquent predisposition</td>
<td>Scale Z—Debasement</td>
</tr>
<tr>
<td>Scale 3—Submissive</td>
<td>Scale D—Sexual discomfort</td>
<td>Scale DD—Impulsive propensity</td>
<td></td>
</tr>
<tr>
<td>Scale 4—Dramatizing</td>
<td>Scale E—Peer insecurity</td>
<td>Scale EE—Anxious feelings</td>
<td>Other</td>
</tr>
<tr>
<td>Scale 5—Egotistic</td>
<td>Scale F—Social insensitivity</td>
<td>FF—Depressive affect</td>
<td>Scale W—Reliability</td>
</tr>
<tr>
<td>Scale 6A—Unruly Scale</td>
<td>G—Family discord</td>
<td>Scale GG—Suicidal tendency</td>
<td></td>
</tr>
<tr>
<td>Scale 6B—Forceful</td>
<td>Scale H—Childhood abuse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale 7—Conforming</td>
<td>Scale 8A—Self-Demeaning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale 8B—Oppositional</td>
<td>Scale 9—Borderline tendency</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


References:


Personality Inventory for Children—Second Edition (PIC-2)
The PIC-2 (Lachar & Gruber, 2001) is a multidimensional clinical measure of behavioral, emotional, and cognitive status for children ages 3–16 years. It is a screening instrument that is usually completed by the parent. The PIC-2 has 275 items in its standard format and contains 12 psychological scales with various subscales. The PIC-2 also contains an abbreviated behavioral summary of 96 items. The psychological scales include Cognitive Impairment, Impulsivity and Distractibility, Delinquency, Family Dysfunction, Reality Distortion, Somatic Concern, Psychological Discomfort, Social Withdrawal, Social Skills Deficits, as well as three Response Validity scales. Parents are asked to respond to the items with True or False answers. The standardization sample generally conformed to U.S. population demographics with the exception of an overrepresentation of Whites and underrepresentation of Hispanics. There was also an overrepresentation of biological parents and an underrepresentation of single parents (Erford & McKechnie, 2006).

No overall composite score is derived, but there are three separate composite scale scores: Externalization-Composite, Internalization-Composite, and Social Adjustment Composite. Raw scores can be converted to T scores when the Student Behavior Survey, a profile form, is completed. Test-retest reliability coefficients ranged from $r = 0.82$ to $0.92$ and internal consistency coefficients ranged from $r = 0.81$ to $0.92$ for the interpreted scales. Criterion validity studies were conducted but did not use other commonly used instruments (Erford & McKechnie, 2006). However, because this new version of the PIC-2 is a major revision of the original, clinicians should be cautious in making diagnostic decisions using the PIC-2 until further research and diagnostic validity studies have been conducted. The PIC-2’s primary benefit continues to be the assessment of parental perceptions of childhood behavioral and clinical difficulties.

References:

Devereux Scales of Mental Disorders (DSMD)
The DSMD (Naglieri, LeBuffe, & Pfeiffer, 1996) is used to assess behaviors related to psychopathology. It can be administered both to individuals as well as groups of children ages 5–18 years in about 15 minutes. There are two forms of the DSMD, the child form and the adolescent form, and each can be rated by parents, teachers, and other appropriate professionals. There are 110 items on this inventory, which measures nine constructs, including Conduct, Attention-Delinquency, Anxiety, Depression, Autism, Acute Problems, Internalizing Composite, Externalizing Composite, and the Critical Pathology Composite. Responses are based on a 5-point scale ranging from Never to Very Frequently. Raw scores can be converted into T scores and percentile ranks. Standardization samples generally conformed to U.S. population demographics for both children and adolescents (Cooper, 2001).

Alpha coefficients were reported at about $r = 0.90$ or higher, and test-retest reliability coefficients were in the 0.80s and 0.90s. Interrater reliability coefficients between parents and teachers were in the 0.40s and 0.50s. This is not surprising given that teachers and parents observe the child’s behavior in two distinct ecological contexts (i.e., school and home). Validity studies yielded adequate results on all levels, with items showing a strong congruence to DSM-IV criteria for the specific behavior disorders in question (Peterson, 2001). There is some dispute in the composition of types of participants used in the reliability and validity study samples and as to whether the type of subjects might have caused elevated coefficients. Even so, there is substantial normative data for the DSMD, and it has emerged as a good assessment for certain antisocial behaviors in children and adolescents.

References:


Children’s Depression Inventory (CDI)

The CDI (Kovacs, 1992) is a self-report inventory used to assess children’s depression. Parent and teacher versions are also available. It can be administered both individually as well as to small groups of children ages 8–17 years in about 10 to 15 minutes. This assessment contains 27 items that cover all nine symptoms for a major depressive syndrome in children as presented in the DSM-III-R. Children’s responses are based on a 3-point scale, from 0 to 2, with 2 being the most severe (Kavan, 1992). Limited normative data are available for the CDI because it was not nationally standardized. The standardization sample was inadequately small and geographically restricted (Knoff, 1992). Scoring was simple and convenient, using the QuickScore™ forms.

Reliability and validity data are also questionable. Although coefficient alphas from two different samples reported in the manual were consistent at \( r = 0.86 \) and 0.87, respectively, many empirical studies yielded inconsistent results. Item–total score coefficients ranged from \( r = 0.08 \) to 0.62. A one-month test-retest reliability coefficient was \( r = 0.43 \), while a nine-week test-retest reliability coefficient was \( r = 0.84 \). Regarding validity, the CDI had adequate correlations with the Revised Children’s Manifest Anxiety Scale but yielded low correlations with Coopersmith Self-Esteem Inventory (Kavan, 1992). The CDI has demonstrated good discrimination between clinical and nonclinical groups (Carey, Gresham, Ruggerio, Faulstich, & Engart, 1987; Hodges, 1990). It is obvious that more empirical data need to be collected with regard to the CDI and it should not be used as a diagnostic tool (Craighead, Curry, & Ilardi, 1995; Fristad, Emery, & Beck, 1997; Knoff, 1992). Admittedly, the construct of depression is more difficult to accurately assess in children than adults because depressive symptoms are more transient in younger clients. In spite of this, the CDI is easy to administer and score and may be helpful during initial clinical assessment (Kavan, 1992). It is, perhaps, the most commonly used screening tool for childhood depression (Craighead et al., 1995; Fristad et al., 1997).

References:


Reynolds Adolescent Depression Scale—Second Edition (RADS-2)

The Reynolds Adolescent Depression Scale—Second Edition (RADS-2) (Reynolds, 2002) is a 30-item self-report inventory for adolescents ages 11–20 years and is designed to assess symptoms associated with depression. Items measure four subscales: Dysphoric Mood (DM, 8 items); Anhedonia/Negative Affect (AN, 7 items); Negative Self-Evaluation (NS, 8 items); and Somatic Complaints (SC, 7 items). Sample items include “I feel lonely,” “I feel like running away,” and “I feel like nothing I do helps anymore.” The items are scored on a 4-point Likert scale (Almost Never, Hardly Ever, Sometimes, or Most of the Time) (Blair, 2005). The RADS-2 is a Level B test and takes about 10 minutes to administer, score, and interpret. The normative restandardization sample (n = 3,300) for the RADS-2 was comprised of an equal number of adolescent males and females living in the United States and Canada. Compared to the 2000 U.S. Census, this sample was considered ethnically diverse and heterogeneous in socioeconomic composition (Reynolds, 2002).

Raw scores are summed to derive a Depression Total score. The Depression Total and four subscales can be converted to a T score or percentile rank according to gender, age group, and gender by age group norms. More than 20 years of research supports the psychometric qualities of the RADS-2, and the new version is found to continue the tradition of a sound instrument (Blair, 2005). Internal consistency of the Depression Total score was $r = 0.92$ (Reynolds, 2002). Test-retest reliability (two weeks) was $r = 0.86$ for the Depression Total score (Reynolds, 2002). Criterion-related validity studies resulted in moderate to high correlations with other measures of depression and indicated the RADS-2 is best used as a screening level test for depression (Erford, 2006). Overall, “the RADS-2 is cost- and time-efficient, easy to use, and a reliable and valid screening instrument for adolescents with symptoms of depression” (Erford, 2006, p. 58).

The RADS-2 is one of the only depression screening tests validated for use with adolescents (Brooks & Kutcher, 2001), and its recommended clinical cutoff of $T = 61+$ has been shown to identify clinically severe symptoms of depression on the Hamilton Depression Rating Scale (HDRS) (Reynolds & Mazza, 1998). The RADS-2 is a screening test and should not be used to supplant use of a clinical interview (Davis, 1990) and is not a substitute for an interview of suicidal ideation (Reynolds, 2002). Volpe and DuPaul (2001) also indicated the RADS-2 shows some usefulness in monitoring the effects of treatment and as one component in a comprehensive diagnostic approach for depression.

References:


Symptom Checklist-90—Revised (SCL-90-R)
The SCL-90-R (Derogatis, 1992) portrays patterns of psychological symptoms in patients and nonpatients. The SCL-90-R can be administered to groups or individuals ages 13 years to adult in about 15 to 20 minutes. Symptoms are measured on 12 constructs: Somatization, Obsessive-Compulsive, Interpersonal Sensitivity, Depression, Anxiety, Hostility, Phobic Anxiety, Paranoid Ideation, Psychoticism, Global Severity Index, Positive Symptom Distress Index, and Positive Symptom Total. There are a total of 90 items on this inventory. Clients are asked to rate their level of discomfort with a particular problem 0 (Not at all) to 4 (Extremely). Norms were constructed on several standardization samples, including psychiatric outpatients, psychiatric inpatients, adult nonpatients, and adolescent nonpatients (Pauker, 1985).

Pauker (1985) and Payne (1985) asserted that the original SCL-90 manual reported satisfactory results for internal consistency ($r = 0.77–0.90$) and test-retest reliability coefficients ($r = 0.78–0.90$, one week apart). The few validity studies conducted portrayed comparable levels to other self-report inventories; however, more research is needed in this area. Other criticisms included a lack of clarity in the manual and the possible limitations inherent in requiring an 8th-grade reading level when using an inventory with adolescents ages 13 years and older. Strengths of the SCL-90-R are the quick administration and scoring procedures as well as its straightforward scoring criteria.

References:
**Millon Index of Personality Styles Revised (MIPS Revised)**

The Millon Index of Personality Styles Revised (MIPS Revised) (Millon, 2003) is a 180-item true-false, Level B, self-report instrument for adults ages 18 years and older and is designed to measure personality styles of normally functioning adults. Scale names and the profile display of the original MIPS were updated to provide administrators with a better, more intuitive approach to interpreting test results. This inventory measures three dimensions of normal personality using 6 Motivating Style scales (Pleasure-Enhancing, Pain-Avoiding, Actively Modifying, Passively Accommodating, Self-Indulging, Other-Nurturing); 8 Thinking Style scales (Externally Focused, Internally Focused, Realistic/Sensing, Imaginative/Intuitive, Thought-Guided, Feeling-Guided, Conservation-Seeking, Innovation-Seeking); 10 Behaving Style scales (Asocial/Withdrawing, Gregarious/Outgoing, Anxious/Hesitating, Confident/Asserting, Unconventional/Dissenting, Dutiful/Conforming, Submissive/Yielding, Dominant/Controlling, Dissatisfied/Complaining, Cooperative/Agreeing); and 4 Validity Indices that provide information about Positive Impression, Negative Impression, Consistency, and Clinical Index. The MIPS Revised takes about 30 minutes to complete using either the paper-and-pencil or computer format. An 8th-grade reading level is required, and it is important to designate age and gender to obtain an accurate report. The MIPS Revised can be scored by hand, computer, mail-in, or optical scanning methods.

The MIPS Revised test offers separate norms for adults and college students, and for both separate and combined genders. The adult sample consisted of 1,000 individuals (500 females, 500 males) ages 18–65 years and is stratified according to the U.S. population by age, race or ethnicity, and education level (Millon, 2003). The college sample consisted of 1,600 students (800 males, 800 females) selected from 14 colleges and universities to be representative of a college student population in terms of ethnicity, age, year in school, major area of study, region of the country, and type of institution. The MIPS Revised can be used as a screening tool in employee selection; for employee assistance programs and leadership and employee development programs; in career planning for high school and college students; in the curriculum for college courses in psychological testing; and in relationship, premarital, marriage, and individual counseling.

**References:**

**Personality Assessment Inventory (PAI)**

The Personality Assessment Inventory (PAI) (Morey, 1991) is a level B test used to assess behaviors related to psychopathology as well as to provide information for screening, clinical diagnosis, and treatment. It can be administered in individual or group formats to clients ages 18 years to adult in about 40 to 50 minutes. There are 344 items on this self-reported inventory, and responses are based on a 4-point scale (Not at All True, Slightly True, Mainly True, and Very True). The PAI requires a 4th-grade reading level. There are 22 nonoverlapping scales, including 4 validity scales (Inconsistency, Infrequency, Negative Impression, Positive Impression); 11 clinical scales (Somatic Complaints, Anxiety, Anxiety-Related Disorders, Depression, Mania, Paranoia, Schizophrenia, Borderline Features, Antisocial Features, Alcohol Problems, Drug Problems); 5 treatment scales (Aggression, Suicidal Ideation, Stress, Nonsupport, Treatment Rejection); and 2 interpersonal scales (Dominance, Warmth). Answers can be scored by hand or by optical scanning, and raw scores can be converted into T scores (Boyle, 1995).

Standardization samples conformed to U.S. population demographics with respect to the test’s diagnostic groups (Kavan, 1995). Reliability of scores seems questionable based on the wide range of coefficients for different variables. Internal consistency coefficients for the 22 scales ranged from $\alpha = 0.45$ to $0.90$, with a median of 0.81 (normative sample); from $\alpha = 0.22$ to 0.89, with a median of 0.82 (college sample); and from $\alpha = 0.23$ to 0.94, with a median of 0.86 (clinical sample). Median alphas were consistent between various races, ages, and genders in the mid to high 0.70s. Test-retest reliability coefficients (3- to 4-week interval) ranged from $r = 0.31$ to $r = 0.92$, with a median of 0.82 (Boyle, 1995). Correlation studies with the Minnesota Multiphasic Personality Inventory (MMPI) and the Marlowe-Crowne Social Desirability Scale yielded mixed validity results. Even with the disputed reliability and validity information, Kavan (1995) viewed the PAI as a competitor of the MMPI-2 that is easier to administer, score, and interpret.

**References:**


California Psychological Inventory (CPI)

The California Psychological Inventory (CPI) (Gough & Bradley, 1996) is a 434-item, level B inventory designed to assess personality characteristics and to predict what people will say and do in specified contexts. The CPI has numerous questions that overlap with the original MMPI but was designed for a different population and purpose than the MMPI (i.e., personality descriptions of a nonclinical population). Scale items measure 20 Folk scales (Dominance, Capacity for Status, Sociability, Social Presence, Self-Acceptance, Independence, Empathy, Responsibility, Socialization, Self-Control, Good Impression, Communality, Well-Being, Tolerance, Achievement via Conformity, Achievement via Independence, Intellectual Efficiency, Psychological-Mindedness, Flexibility, and Femininity-Masculinity); 3 Vector scales (Internality-Externality, Norm-Questioning-Favoring, and Self-Realization); and 13 Special Purpose scales. These scales are for clients ages 13 years and older, are written at a 5th-grade reading level, and take about 45 to 60 minutes to administer (Atkinson, 2003). The CPI is self-administered and can be done using either pencil and paper or a computer. Forms are scanned for automated data entry. Using the scores from the three Vector scales, a cuboidal personality typology is developed, which helps to classify individuals into four categories (Atkinson, 2003).

The normative sample ($n = 6,000$; 3,000 of each gender) was reported as not being representative or random because of use of primarily high school students (50%) and undergraduate students (16.7%), so these are probably the best populations for which to use the instrument, though the manual provides useful reference tables for comparing students of various ages (Hattrup, 2003). The test produced internal consistency Cronbach’s alpha estimates on the 20 Folk scales ranging from $\alpha = 0.43$ to 0.85, with a median of 0.76. For the three Vector scales, the internal consistency estimates ranged from $\alpha = 0.77$ to 0.88. Cronbach’s alpha for the 13 specialty scales ranged from $\alpha = 0.45$ to 0.88. Alpha reliabilities of the CPI scales ranged from $\alpha = 0.62$ to 0.84 in the total sample, with a median of 0.77. Test-retest reliabilities were based on samples of 108 males and 129 females who were retested after a 1-year interval, and samples of 91 females and 44 males who were retested after 5- and 25-year intervals, respectively. For the 1-year retest, scale reliabilities ranged from $r = 0.51$ to $r = 0.84$, with a median of 0.68. For the 5-year and 25-year retest, reliabilities ranged from $r = 0.36$ to $r = 0.73$, and $r = 0.37$ to $r = 0.84$, respectively. Test-retest reliability estimates among high school students were between 0.60 and 0.80 for a 1-year period. The Folk and Vector scales had moderate to strong construct validity correlation scores (0.40 to 0.80), but the predictive power regarding individual behavior in a given situation was weak.

References:


Tennessee Self-Concept Scale—Second Edition (TSCS-2)

The Tennessee Self-Concept Scale—Second Edition (TSCS-2) (Fitts & Warren, 1996) is one of the most commonly used self-report measures of self-concept and can be used for children and adults. The test was standardized on 3,000 subjects, ages 7–90 years, and can be administered to individuals or groups in about 10 to 20 minutes. The Adult Form is designed for clients ages 13 years or older and has 82 items. The Child Form is designed for students ages 7–14 years and has 76 items. A Short Form consisting of the first 20 items of either form can be used as well. Items comprise 15 subscales and a total Self-Concept score (see Table). The items are rated on a 5-point Likert scale ranging from Always False to Always True. The TSCS-2 can be hand-scored in approximately 10 minutes, or computer-scored. Reliability is adequate, with lower internal consistencies on subscales than Total Self-Concept, ranging from $\alpha = 0.73$ to 0.93. Test-retest reliability scores ranged from $r = 0.47$ to $r = 0.83$ (Brown, 1998). Fitts and Warren (1996) reported acceptable levels of score validity for the TSCS-2.

Table. Scales on the Tennessee Self-Concept Scale—Second Edition

<table>
<thead>
<tr>
<th>Self-concept scores</th>
<th>Supplementary scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical</td>
<td>Identity</td>
</tr>
<tr>
<td>Moral</td>
<td>Satisfaction</td>
</tr>
<tr>
<td>Personal</td>
<td>Behavior</td>
</tr>
<tr>
<td>Family</td>
<td>Validity scores</td>
</tr>
<tr>
<td>Social</td>
<td>Inconsistent</td>
</tr>
<tr>
<td>Academic/Work</td>
<td>Responding</td>
</tr>
<tr>
<td><strong>Summary scores</strong></td>
<td>Self-criticism</td>
</tr>
<tr>
<td>Total self-concept</td>
<td>Faking good</td>
</tr>
<tr>
<td>Conflict</td>
<td>Response distribution</td>
</tr>
</tbody>
</table>


References:

Rorschach Inkblot Test

The Rorschach Inkblot Test (Rorschach, 1921/1998), originally developed by Hermann Rorschach in 1921, is the best-known and most used projective test. The test’s purpose is to assess how a client perceives and organizes thoughts about the world. The test is a Level C instrument and is individually administered to clients ages 5 and older, in about 20 to 30 minutes (Hess, Zachar, & Kramer, 2001). It consists of 10 plates of bilaterally symmetrical inkblots (Janda, 1998): 5 are black and white; 2 are black, white, and red; and the remaining 3 are comprised of pastel colors (Hess et al., 2001). Clients are presented with the cards and asked what they think of the inkblot or what it might be. In the second part of administration, clients are asked to explain their original answers. Scoring and interpretation are frequently completed using a scoring system originally developed by John Exner in the 1970s called the Comprehensive System for Administering, Scoring, and Interpreting the Rorschach (Exner, 2002). Exner’s multifaceted system involves interpretation of three aspects of responses: Location (W for the entire blot, D major portion of the blot, and Dd for uncommon responses); Determinants (there are nearly two dozen having to do with shape, activity of humans, chromatic features, etc.); and Content (there are 26 categories used to interpret the content of the story). A Structural Summary is composed based on an interpretive rating scale developed by Exner (Janda, 1998).

As with many projective tests, it is often hard to find concrete empirical data on the Rorschach. Subjectivity is such a part of interpretation, and there can be definite diversity in administration procedures depending on testing purpose and clinician training. It has been noted that well-trained users of Exner’s scoring system agree on the major variables over 88% of the time (Hess et al., 2001). Still, there is substantial debate over the interrater reliability of Exner’s system. Exner purports that test-retest reliability estimates are at or above $r = 0.70$ at both 1-year and 3-year intervals. According to Hess et al. (2001), validity data of the Rorschach also yield many questions and concerns. Various questions of subjectivity arise based on administration, scoring, and interpretation procedures. Still even with the lack of standardization and empirical data, the Rorschach used in conjunction with Exner’s Comprehensive System (2002) is a better personality test than most opponents will acknowledge (Hess et al.). Critics of the Rorschach point out that statistical prediction is usually more accurate than clinical prediction (i.e., judgment), and the Rorschach relies primarily on clinical prediction to measure personality. Far more psychometric research needs to be done using the Rorschach, but it has the potential to generate meaningful personality data (Hess et al., 2001).

References:


Coping Inventory for Stressful Situations (CISS)

The Coping Inventory for Stressful Situations (CISS) is a 48-item self-report inventory used to assess three major coping styles: (1) task-oriented, (2) emotion-oriented, and (3) avoidance-oriented. Each coping style is assessed through 16 items. The CISS is based on Endler’s (one of the authors of the CISS) multidimensional interaction model of stress, anxiety, and coping. According to Endler (1997), task-oriented coping contains efforts such as problem solving and situation changing, whereas emotion-oriented coping contains self-oriented responses such as emotional reactions, self-preoccupation, and fantasizing. Avoidance-oriented coping contains activities or cognitive changes to avoid stressful situations (for details of the multidimensional interaction model and the three coping styles, see Endler, 1997). There are two versions of the CISS: an Adolescent version (ages 13–18) and an Adult version (ages 18 and older). Paper-and-pencil record forms called “QuikScore™” are available. A 21-item brief format for adults (CISS: Situation Specific Coping [CISS:SSC]) is also available to assess coping style in situations involving social evaluation and interpersonal conflicts (Multi-Health Systems, Inc., 2003). Current cost information and online ordering information are available on the website of Multi-Health Systems (2003).

Each item of the CISS is formatted on a Likert scale ranging from 1 (Not at All) to 5 (Very Much). The CISS takes approximately 10 minutes to complete and has a Level A qualification for administration and interpretation. An examiner scores the CISS using a scoring grid and obtains a percentile rank and T score using a profile sheet on the back side of the scoring grid. Provided scales are Task, Emotion, and Avoidance. Avoidance consists of two subscales: Distraction and Social Diversion.

Norms are provided for adults and adolescents (Tirre, 2003). For adults, separate male and female norms are provided for general-population and psychiatric patients, respectively. For adolescents, separate norms are provided for individuals ages 13–15 years and 16–18 years. Separate college student norms are also available. Endler (1997) found sufficient internal consistency and test-retest reliability for the CISS. Endler (1997) also found the scores of the CISS to be valid. Through an examination of construct validity, Endler discovered that some CISS scales were significantly correlated with related measures, such as the Beck Depression Inventory (BDI) and the Eysenck Personality Inventory (EPI).

Professional counselors interested in using the CISS are encouraged to explore Endler’s multidimensional interaction model prior to use. Endler (1997) insisted on the necessity of examining not only the interaction between person and situation variables, but also the interaction within person variables (e.g., cognitive style, biological variables) and situation variables (e.g., stressful events, physical environments), given that “stress, anxiety, and coping all involve complex processes and all interact with one another” (Endler, 1997, p. 149).

References:
Metropolitan Achievement Test—Eighth Edition (Metropolitan-8)
The Metropolitan Achievement Test—Eighth Edition (Metropolitan-8) (Harcourt Educational Measurement, 2001) consists of 13 levels of tests designed for grades kindergarten (Preprimer) through the junior or senior year in high school (Secondary 3). The test is available in a Complete Battery (Reading, Math, Language, Science, and Social Studies) or a Short Form (Reading Comprehension, Math Concepts & Problem Solving, and Language). A separate test of Writing is also included, as well as open-ended response versions for Reading and Math. Each subtest contains between 20 to 50 items, depending on the level of the test, and administration time limits are adjusted accordingly. Total administration time may take from 1 1/2 to over 5 hours, depending on the grade level and test given. Most items are multiple-choice, with the exception of the open-ended items for Reading (e.g., rubrics) and Math. Open-response items are scored on a scale of 0 to 3, although no scoring criteria are given ahead of administration to help prepare students to craft acceptable responses (Lukin, 2005). Scores are used to place students into the performance standard categories of Advanced, Proficient, Basic, and Below Basic (Harwell, 2005).

Norms used for the Metropolitan-8 generally conformed to U.S. population characteristics, with the critical exception of an overrepresentation of rural schools (51.3%, compared to 29.6% nationally) and an underrepresentation of urban classrooms (10.1%, versus 31.9% nationally). This error in sample representativeness urges caution in interpreting scores for students in urban settings. Internal consistency (KR-20) reliabilities were generally quite high (most exceeding $r = 0.90$), with the exception of the Science and Social Studies subtests in the earlier grades (Harwell, 2005). The majority of test-retest reliabilities were in the $r = 0.70–0.90$ range, with the lower values again associated with the Science and Social Studies subtests. Regarding validity, the test author’s primary consideration was a matching of the Metropolitan-8 to the local school curriculum. This information does not provide compelling evidence that the test scores are valid for their intended use (Lukin, 2005).

References:


The Stanford Achievement Test—Tenth Edition (Stanford-10) (Harcourt Brace Educational Measurement, 2003) was designed to measure student progress toward high academic standards. It is used in school districts all over the country as a way to inform educators of what their students know and what they are able to achieve. The Stanford-10 tests the areas of Reading, Mathematics, Language, Spelling, Listening, Science, and Social Studies, from kindergarten through senior year of high school. The Stanford-10 is an untimed test, but testing guidelines are given to administrators to allow for proper planning (The Psychological Corporation, 2005). The Stanford-10 has two equivalent forms (Forms A and B).

All items on the Stanford-10 are multiple-choice. Raw scores are converted to scaled scores, national and local percentile ranks, stanines, grade equivalents, and normal-curve equivalents. The normative sample generally conformed to 2002 U.S. population demographics (The Psychological Corporation, 2005). KR-20 reliability coefficients for 8th-grade students ranged from $r = 0.71$ to $r = 0.92$ (median = 0.86) for Form A, and from $r = 0.70$ to $r = 0.91$ (median = 0.86) for Form B. Alternate-form reliability coefficients for subtests ranged from $r = 0.69$ to $r = 0.85$ when comparing Form A to Form B (Harcourt Assessment Inc., 2004a, 2004b).

References:


Metropolitan Readiness Tests—Sixth Edition (MRT-6)
The Metropolitan Readiness Tests—Sixth Edition (MRT-6) (Nurss & McGauvran, 1995) was designed to assess basic and advanced skills necessary to begin reading and mathematics. It is composed of two levels: Level 1 is individually administered and was developed to assess the evolving literacy and mathematic strategies in pre-kindergarten and beginning kindergarten students; Level 2 is group administered and assesses beginning reading and mathematics strategies and processes (Novak, 2001). Level 1 is composed of six subtests: Auditory Memory, Rhyming, Letter Recognition, Visual Matching, School Language & Listening, and Quantitative Language. Level 2 contains eight subtests: Beginning Consonants, Sound-Letter Correspondence, Visual Matching, Finding Patterns, School Language, Listening, Quantitative Concepts, and Quantitative Operations (Cohen & Swerdlik, 2010). Administration of each level of the MRT-6 is broken down into four sittings and requires approximately 85 minutes total. Raw scores are converted to percentile ranks, normal-curve equivalents, scaled scores, and standard scores. The standardization sample generally conformed to U.S. population demographics (Novak, 2001).

The MRT-6 provides information that may be useful in determining early readiness skills in reading and math. However, it should not be used for analysis of individual strengths and weaknesses. Internal consistency estimates determined that Quantitative Concepts and Reasoning subtests may be used for screening purposes. The estimates of the remaining subtests were below $r = 0.80$ and therefore insufficient for decision making regarding individual students. The Story Comprehension subtest was the weakest overall, with coefficients ranging from $r = 0.53$ to $r = 0.77$ (Novak, 2001). There is limited evidence of validity, and what does exist does not support the MRT-6’s routine use for the screening of academic achievement (Kamphaus, 2001).

References:


Peabody Individual Achievement Test-Revised (PIAT-R)

The Peabody Individual Achievement Test-Revised (PIAT-R) (Markwardt, 1998) is a survey of scholastic attainment for students ages 5 to 18 years and takes about 60 minutes to administer. The test is composed of six subtests, each with 100 possible raw score points or items: General Information, Reading Recognition, Reading Comprehension, Spelling, Mathematics, and Written Expression. Some PIAT-R items follow a multiple-choice format, others a free-response format. Items on each subtest are arranged in order of increasing difficulty. The basal score is established for each subtest through five consecutive correct responses, and the ceiling is composed of seven consecutive responses that contain five errors (because of the multiple-choice format). Raw scores are converted into age- and grade-based standard scores, age and grade equivalents, percentile ranks, and stanines. The standardization sample generally conforms to the U.S. population estimates (Markwardt, 1998).

According to the manual (Markwardt, 1998), test-retest reliability for subtests and the total battery for selected grades ranged from $r = 0.67$ to $r = 0.98$ (median = 0.91). Internal consistency coefficients ranged from $r = 0.87$ to $r = 0.98$ (median = 0.95). The large number of items in each subtest led to very stable score reliability. However, this may be because all items below the basal level are counted as correct, and all items above the ceiling level are considered incorrect (Cross, 2001). Importantly, the Written Expression subtest displayed an extremely small range of scores and poor psychometrics; thus norm-referenced scores were not derived for this subtest. Construct validity is demonstrated through correlations with the original PIAT and the Peabody Picture Vocabulary Test—Revised (PIAT-R) (Dunn & Dunn, 1997). The PIAT-R is a well-written test constructed to give a greater understanding of an individual’s achievement, as well as to pinpoint the examinee’s strengths and weaknesses in basic academic areas (Cross, 2001).

References:


Kaufman Test of Educational Achievement—Second Edition (KTEA-II)

The Kaufman Test of Educational Achievement—Second Edition (KTEA-II) (Kaufman & Kaufman, 2005) is used in the screening and diagnosis of academic achievement for children in grades 1–12, or ages 6–22 years. It is available in two formats: brief and comprehensive. The Brief Form contains three subtests: Reading (word recognition and reading comprehension); Mathematics (computation and application problems); and Written Expression (written language and spelling). The Comprehensive Form contains six components, each with two or more subtests: Reading Composite (Letter & Word Recognition and Reading Comprehension); Reading-Related Subjects (Phonological Awareness, Nonsense Word Decoding, Word Recognition Fluency, Decoding Fluency, Associational Fluency, and Naming Facility); Math Composite (Math Concepts & Applications and Math Computation); Written Language Composite (Written Expression and Spelling); Oral Language Composite (Listening Comprehension and Oral Expression); and the Comprehensive Achievement Composite. The grade level of the individual determines the starting point for each subtest, and testing stops (i.e., ceiling) if the student fails every item in one unit of the subtest. Depending on a client’s grade and ability level, the Brief Form may take 20 to 30 minutes to administer, while the Comprehensive Form may take between 30 and 80 minutes. Scoring is dichotomous. Interpretive scores include standard scores ($M = 100; SD = 15$), age and grade equivalents, percentile ranks, stanines, and normal-curve equivalents (Salvia & Ysseldyke, 2004).

The authors of the KTEA-II did an excellent job of describing and documenting the characteristics of the normative sample. It was shown in detail to generally conform to the U.S. population (Salvia & Ysseldyke, 2004). While Venn (1994) reported that the original K-TEA was well designed and provided a complete system for measuring academic achievement, a measure of written expression was noticeably absent. This was corrected in the second edition. Internal consistency and test-retest reliability for scores on subtests were acceptable for screening-level purposes. Both forms showed evidence of content validity (Kaufman & Kaufman, 2005). Miller (1999) reported that detailed procedures for error analyses in the previous K-TEA allowed examiners to compare a student’s errors to that of grade mates. However, he cautioned that “neither the brief or comprehensive forms appear to have a sufficient number of items at the level of emerging academic abilities to discriminate between deficient and low-normal achievements” (p. 153). New research on the KTEA-II will determine whether scale improvement will address this concern.

References:


Wide-Range Achievement Test—Fourth Edition (WRAT-4)
The Wide-Range Achievement Test—Fourth Edition WRAT-4 (Wilkinson & Robertson, 2006) is a brief (15- to 35-minute for 5-7 year olds; 35-45 minutes for 8-94 year olds) screening test of achievement with four subtests: Word Reading, Sentence Comprehension, Math Computation, and Spelling. The Word Reading and Sentence Comprehension subtests must be administered individually, but the Math Computation and Spelling subtests can be administered to individuals or groups. Alternate forms (Blue Form and Green Form) are available, and the test can be used with clients aged 5–94 years. The Spelling and Word Reading subtests are untimed, but the Math Computation subtest has a 15-minute time limit. Items are scored correct (1 point) or incorrect (0 points). The “5/10 rule” is used to determine basal (5 consecutive correct items) and ceiling levels (10 consecutive incorrect items). The stratified standardization sample generally conformed to U.S. population demographics based on the 2000 U.S. Census.

The WRAT-4 is an inexpensive, easy-to-administer screening test but lacks the comprehensiveness to identify specific learning deficiencies. Raw scores are converted to standard scores ($M = 100; SD = 15$), percentile ranks, and age and grade equivalents. Internal consistency estimates ranged from $r = 0.83$ to 0.89 for the Math Computation subtest; $r = 0.88$ to 0.91 for the Spelling subtest; $r = 0.92$ to 0.93 for the Word Reading subtest, and $r = 0.90$ to 0.93 for the Sentence Comprehension subtest across all age categories (Wilkinson & Robertson, 2006)—very acceptable for a screening-level test. Knoop (2004) pointed to the WRAT-3’s versatility in distinguishing difficulties in the initial stages of basic academic skills, and in vocational or rehabilitation counseling when questions are raised over a client’s ability to function adequately in the workplace. It appears to also be a robust instrument across cultures.

References:

**Essential Skills Screener (ESS)**

The Essential Skills Screener (ESS) (Erford, Vitali, Haas, & Boykin, 1995) was designed to identify 3- to 11-year-old children who may be at risk of educational problems or failures. It is composed of reading, math, and writing screening tests for each of three age levels: preschool, early elementary, and upper elementary (see Table). Each test typically takes less than 10 minutes to administer and score. Dichotomous scoring is used, and raw scores are converted into standard scores, percentile ranks, and age and grade equivalents. The standardization sample is generally representative of the U.S. population (Erford et al., 1995).

Internal consistency for all nine screeners ranged from 0.87 to 0.95 (median = 0.93), and test-retest coefficients ranged from 0.83 to 0.93 (median = 0.90), indicating acceptable levels of reliability for a screening-level purpose. Sufficient evidence of score validity has been amassed (Erford, 1997, 1999, 2004; Erford, Ivey, & Dorman, 1999; Erford, Ivey, Dorman, & Wingeart, 2001; Erford & Stephens, 2005; Erford et al., 1995; Erford et al., 1998).

**References:**


<table>
<thead>
<tr>
<th>Screener</th>
<th>Age/grade of students</th>
<th># items</th>
<th>Facets assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Preschool</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading ESS—Preschool Version (RESS-P)</td>
<td>Ages 3–5</td>
<td>51</td>
<td>Picture Vocabulary, Visual Discrimination, Visual Figure-Ground, Letter Identification, Experience with Books</td>
</tr>
<tr>
<td>Writing ESS—Preschool Version (WESS-P)</td>
<td>Ages 4–5</td>
<td>23</td>
<td>Form Copying, Copying Speed, Letter/Number Copying, Name Writing</td>
</tr>
<tr>
<td><strong>Early Elementary</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading ESS—Elementary Version (RESS-E)</td>
<td>Ages 6–7, Grades 1–3</td>
<td>67</td>
<td>Letter Identification, Consonant Letter-Sound Association, Digraphs and Blends, Sight-Word Vocabulary, Oral Reading/Passage Comprehension</td>
</tr>
<tr>
<td>Writing ESS—Elementary Version (WESS-E)</td>
<td>Ages 6–8, Grades 1–2</td>
<td>13</td>
<td>Name Writing, Writing Speed, Spelling, Sentence Writing</td>
</tr>
<tr>
<td><strong>Upper Elementary</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading ESS—Upper Elementary Version (RESS-U)</td>
<td>Ages 8–11, Grades 4–6</td>
<td>35</td>
<td>Sight-Word Vocabulary, Digraphs and Blends, Oral Reading/Passage Comprehension</td>
</tr>
<tr>
<td>Writing ESS—Upper Elementary Version (WESS-U)</td>
<td>Ages 9–11, Grades 3–6</td>
<td>14</td>
<td>Writing Speed, Spelling, Story Composition</td>
</tr>
</tbody>
</table>

Gray Oral Reading Test—Fourth Edition (GORT-4)
The Gray Oral Reading Test—Fourth Edition (GORT-4) (Weiderholf & Bryant, 2001) was designed to measure growth in oral reading as well as aid in the diagnosis of oral reading difficulties. It measures the rate, accuracy, fluency, and comprehension of student oral reading ability for ages 6-0 to 18-11. The GORT-4 is composed of two parallel forms, each containing 14 separate stories with five multiple-choice comprehension questions for each story. It takes 20 to 30 minutes to administer and score. The norms used are considered representative of the current U.S. population (Crompton, 2003).

According to Crompton (2003), the revision of the GORT-4 has led to an edition that meets technical challenges. It provides the test user with accurate information reflecting the current theoretical rationale in measuring reading ability. However, the examiner should have considerable practice prior to administering the GORT-4 to be able to sufficiently prompt the reader, time the reader’s rate, and mark deviations from the print while testing. Reliability was sufficient, with test-retest scores ranging from $r = 0.85$ to $r = 0.95$. Coefficient alphas from 13 age groups averaged from $r = 0.91$ to $r = 0.97$. Equivalence of parallel forms was documented (Crompton, 2003).

References:

Woodcock Reading Mastery Tests—Third Edition (WRMT-III)
The nine subtests of the Woodcock Reading Mastery Tests—Third Edition (WRMT-III) (Woodcock, 2011) (Phonological Awareness, Listening Comprehension, Letter Identification, Word Identification, Rapid Automatic Naming, Oral Reading Fluency, Word Attack, Word Comprehension, and Passage Comprehension) combine to measure several important aspects of reading ability. The WRMT-III was designed to test individuals ages 4.5–79 years and requires about 15-45 minutes to administer. A short scale, consisting of the Word Identification and Passage Comprehension subtests, may also be given in about 15 minutes. The scoring is dichotomous, and raw scores can be converted to percentile ranks, standard scores, age and grade equivalents, and instructional ranges (Crocker, 2001).

The WRMT-III shows good technical quality and diagnostic value. The standardization sample included 3,360 individuals aged 4.5 to 79 years. Internal consistency coefficients ranged from $r = 0.68$ to $r = 0.98$ (median = 0.91). Split-half reliability estimates ranged from $r = 0.86$ to $r = 0.99$ (median = 0.97). Content and concurrent validity were shown, as well as intercorrelations with other tests (Woodcock-Johnson Tests of Achievement—Third [WJ-III], Wechsler Individual Achievement Test – Third Edition [WIAT-III], Kaufman Tests of Educational Achievement – Second Edition [KTEA-II]) (Pearson, 2012).

References:


**Stanford Diagnostic Reading Test—Fourth Edition (SDRT-4)**

The Stanford Diagnostic Reading Test—Fourth Edition (SDRT-4) (Karlsen & Gardner, 1996) helps to determine a student’s strengths and weaknesses in the major skills of reading. It can also be used to find patterns of reading skills in schools or districts (Engelhard, 1998). The main categories tested are Decoding or Phonetic Analysis, Vocabulary, Comprehension, and the Rate of Reading or Scanning. There are six levels of the test, and the upper three levels have two alternate forms. These levels can be used for students from 1st grade through the first semester of college (Swerdlik & Bucy, 1998).

Norm-referenced and criterion-referenced information is provided in the manual, in addition to a detailed and comprehensive explanation of the standardization sample (Engelhard). Raw scores can be converted into progress indicators, which help to identify students who have or have not demonstrated basic proficiency in certain skills important to the developmental sequence of reading, as well as percentile ranks, stanines, grade equivalents, and scaled scores using the tables provided in the manual. The instructions for the examiner are clear and detailed, and the teacher’s manual gives some helpful information and tips for teaching reading. The alternate-form reliability ranged from $r = 0.62$ to $r = 0.82$, and content validity, criterion-related validity, and construct validity are all presented. Overall, the psychometric results are sufficient, but low enough on some subtests to rule out using the SDRT-4 for individual diagnosis (Engelhard).

**References:**


Stanford Diagnostic Mathematics Test—Fourth Edition (SDMT-4)

The Stanford Diagnostic Mathematics Test—Fourth Edition (SDMT-4) (Beatty, Madden, Gardner & Karlsen, 1996) is a level B test designed to measure a student’s mastery of basic math skills and concepts. The purpose of the SDMT is to identify area of math deficiency so clients can be given the help they need on an individualized basis. The test can be given to individuals or to a small group. Nagy (1998) indicated that the test tries to be diagnostic and normative at the same time but does not have great success doing either; although it is probably better used diagnostically. The test consists of six levels: (1) Red (grades 1.5–2.5); (2) Orange (grades 2.5–3.5); (3) Green (grades 3.5–4.5); (4) Purple (grades 4.5–6.5); (5) Brown (grades 6.5–8.9); and (6) Blue (grades 9–13). Each of the three lower levels has one form, and the three upper levels have two alternate forms (Lehmann, 1998). On each level, the questions test the student’s understanding of concepts and applications (32 multiple-choice and 30 free-response questions) and computation (20 multiple-choice and 20 free-response questions).

Problem solving and strategies are emphasized (Lehmann, 1998), but the content focused on at each level differs. This test has more low-difficulty questions than most mathematics achievement tests, which makes it especially helpful for assessing the performance of below-average students, because they will have some success. Because this test is designed specifically to find weaknesses, it is less helpful for higher-achieving students. Each level is accompanied by a manual that is well laid out and has clear instructions. The SDMT-4 can be hand-scored or sent to The Psychological Corporation for computer scoring. Three of four reports are given as part of the Basic Scoring Service, and the fourth, the Group Roster Summary, is optional. Raw scores can be converted into progress indicators that help to identify students who have or have not demonstrated basic proficiency in certain skills important to the developmental sequence of reading, percentile ranks, stanines, grade equivalents, and scaled scores using tables provided in the manual. Numeric, graphic, and verbal information can be obtained for the class as well as information on where the class falls according to national norms.

The standardization sample was generally representative of U.S. demographics, and the test items have been well researched and developed. In addition, the content validity is adequate, as is the reliability, although it may indicate that the test is adequate only for screening purposes because of the small numbers of items (Lehmann, 1998).

References:


Cognitive Abilities Test (CogAT)
The Cognitive Abilities Test (CogAT) (Lohman & Hagen, 2001) provides a measurement of level and pattern of cognitive development. It is a level B instrument. The test can be administered to students K–12 and has 11 different levels. Both the primary battery (Levels K, 1, and 2) and the multilevel battery (Levels A–H) are broken into three batteries of skills: Verbal, Quantitative, and Nonverbal. The primary battery takes three sessions of 35 to 40 minutes each; the multilevel battery takes three sessions of 45 to 50 minutes each. Levels K, 1, and 2 have six subtests: Oral Vocabulary, Verbal Reasoning, Relational Concepts, Quantitative Concepts, Figure Classification, and Matrices. Levels A–H (used for 3rd- through 12th-graders) have nine subtests: Sentence Completion, Verbal Classification, Verbal Analogies, Quantitative Relations, Number Series, Equation Building, Figure Classification, Figure Analogies, and Figure Analysis. The number of items varies by level and subtest. Four scores—Verbal, Quantitative, Nonverbal, and Total—are given at each level, but subtest scores are not reported individually. The CogAT was standardized at the same time as the Iowa Tests of Educational Development (ITBD) and the Iowa Test of Basic Skills (ITBS) and conforms to U.S. population demographics (Salvia & Ysseldyke, 2004).

The CogAT offers a wide range of test levels for different-aged students. Standard age scores ($M = 100; SD = 15$), national grade equivalents, age-based percentile ranks, and grade- and age-based stanines are available for the CogAT. Total battery score reliability coefficients for Level K range from $r = 0.85$ to $r = 0.89$; for Level 1, from $r = 0.87$ to $r = 0.93$; for Level 2, from $r = 0.86$ to $r = 0.92$; and for Levels A–H, from $r = 0.93$ to $r = 0.95$. The total-battery range is very high, from $r = 0.94$ to $r = 0.98$. Minimal validity data are provided in the manual.

References:
**Slosson Intelligence Test—Primary (SIT-P)**

The SIT-P (Erford, Vitali, & Slosson, 1999) is a brief, 121-item screening-level test of typical developmental skills of early childhood that is used for children ages 2–7 years. It usually requires about 30 minutes to administer. A basal level is established when a child correctly answers seven consecutive questions; a ceiling is determined when the child is unable to answer seven consecutive items (Erford et al., 1999). Various theoretical models of intelligence were incorporated into the development of the test. Verbal and nonverbal items are used, drawing from many of the domains used in the WISC-III or the SBIS-4. Items are scored correct (1 point) or incorrect (0 points), and are simply totaled to reach the Verbal Scale Total Score, Nonverbal Scale Total Score, and Total Score. Conversion tables in the manual convert the scores to percentile ranks, age equivalencies, and standard scores ($M = 100; SD = 15$) (Erford et al., 1999).

The standardization sample was made up of children from Virginia and Maryland ages 2–7 years. The sample generally conformed to the 1990 U.S. Census statistics, though the socioeconomic level of the sample was slightly higher than average, and there were slightly more White and slightly fewer Hispanic children in the sample. KR-20 coefficients ranged from $r = 0.72$ to $r = 0.97$, and internal consistency coefficients for the Verbal, Nonverbal, and Total scales for the entire standardization sample were $r = 0.95$, 0.97, and 0.98, respectively (Erford, Vitali, & Slosson, 1999). The SIT-P was correlated with scores on similar tests (SBIS-4, WISC-III, SIT-R TSS, PPVT-R, and EOWPVT-R), and the total-scale score correlation ranged from 0.46 to 0.74. Several techniques were used to demonstrate the construct validity of the SIT-P, and all highly asserted construct validity.

**References:**
**Kaufman Brief Intelligence Test—Second Edition (KBIT-2)**

The Kaufman Brief Intelligence Test—Second Edition (KBIT-2) (Kaufman & Kaufman, 2004b) is a quick (15- to 30-minute) estimate of intelligence and can be used to screen students for potential educational problems and gifted programs. It can be used with clients ages 4–90 years. The KBIT-2 consists of an IQ composite score with two subtests: Vocabulary (verbal, crystallized) and Matrices (nonverbal, fluid). Items are scored right (1 point) or wrong (0 points). During the standardization process, adequate consideration was given to U.S. population demographics based on the U.S. Census (Overton, 1996). Raw scores can be converted to standard scores and national percentile ranks by age. Split-half reliability and test-retest coefficients are satisfactory, ranging between 0.80 and 0.90. Criterion-related validity studies were conducted with the original K-BIT, the WISC-R, the K-ABC, and the WAIS-R, yielding satisfactory validity coefficients (Overton, 1996). Hayes (1999) reported the original KBIT appeared culture-fair and relatively free of gender or ethnic bias. Overall, the KBIT-2 is a brief, efficient, and technically adequate screening measure for a balanced estimate of intellectual ability.

**References:**


Kaufman Assessment Battery for Children—Second Edition (KABC-II)
The KABC-II (Kaufman & Kaufman, 2004a) is an individually administered test used to assess cognitive ability. The test is based on two theoretical models, the Luria neuropsychological model (which excludes verbal ability) and the Cattell-Horn-Carroll model (CHC). The test administrator can choose which form, the Luria or the CHC, to use based on the client’s cultural and/or language background. The Luria format takes 25 to 55 minutes to administer, while the CHC format requires 35 to 70 minutes. The KABC-II can be used with clients aged 3–18 years, with applicable subtests varying by age. The five scales (and related subtests) on the KABC-II include Simultaneous (Triangles, Face Recognition, Pattern Reasoning, Block Counting, Story Completion, Conceptual Thinking, Rover, Gestalt Closure); Sequential (Word Order, Number Recall, Hand Movements); Planning (Pattern Reasoning, Story Completion); Learning (Atlantis, Atlantis Delayed, Rebus, Rebus Delayed); and Knowledge (CHC model only) (Riddles, Expression Vocabulary, Verbal Knowledge). The standardization sample of 3,025 individuals generally conformed to U.S. population demographics in terms of ethnicity and was generally equally divided by sex (Kaufman & Kaufman, 2004a).

The KABC-II is a culturally sensitive measure of cognitive ability. Raw scores can be transformed into age-based scaled scores, standard scores (M = 100; SD = 15), age equivalents, and percentile ranks. Test scale scores had mean reliability coefficients ranging from 0.87 to 0.96. Subtest score mean reliability coefficients were more variable, ranging from 0.74 to 0.92. The validity of the original KABC has been widely disputed; however, the validity of the revised KABC-II seems adequate according to the technical manual (Kaufman & Kaufman, 2004a).

Reference:

The WJ-III-NU COG (Woodcock, McGrew & Mather, 2001) is a comprehensive assessment of intelligence and general ability based on the Cattell-Horn-Carroll (CHC) model of cognitive abilities, one of the best and most theoretically sound descriptions of the structure of intelligence (Keith, Kransler, & Flanagan, 2001). It is designed to help with diagnosis of learning disabilities, determine educational discrepancies, develop educational plans for individuals, assess growth, and offer guidance in educational and clinical settings for anyone over 24 months of age. A total of 20 subtests (see Table) can be administered individually (some can be adapted for group administration). Depending on examiner or examinee needs, different combinations of the tests can be used. For example, by using the Verbal Comprehension, Concept Formation, and Visual Matching subtests, a Brief Intelligence Ability score can be derived that is helpful for screenings or reevaluations without a comprehensive assessment. Likewise, a General Intellectual Ability score can also be found using multiple subtests from the battery.

Each subtest is easy to administer and takes about 5 minutes to complete. Scoring must be accomplished using the accompanying Compuscore software. There is a standard and an extended version of the battery. The standard version of the test is made up of the first 10 subtests, and the extended version includes the last 10 subtests. The subtests were designed to measure the seven factors of the Cattell-Horn-Carroll model: Comprehension-Knowledge (Gc); Long-Term Retrieval (Glr); Visual-Spatial (Gv); Reasoning (Gf); Processing Speed (Gs); and Short-Term Memory (Gsm).

This norm-referenced test is very useful for developing a wide range of derived scores, including standard scores ($M = 100; SD = 15$), age and grade equivalents, percentile ranks, discrepancy scores, and many scores reported using the scales developed for this specific test (Cizek, 2003). The reliability and validity results are impressive, and the test’s normative data are sufficient to be able to use the test as an accurate diagnostic tool. Cizek commented that some diagnostic examples and intervention strategies would be useful, in addition to some Compuscore reports and interpretations. Sandoval (2003) commented that all of the materials are very well done and useful, and the manual is very comprehensive. Rizza, McIntosh, & McGunn (2001) indicated the cognitive battery provides multiple means for analyzing specific strengths and weaknesses of gifted students. Proctor and Prevatt (2003) reported that the six different types of discrepancies scores derived can be very helpful in the diagnosis of learning disabilities.

References:


<table>
<thead>
<tr>
<th>Subtest and factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Verbal Comprehension (Gc)</td>
<td>Measures lexical knowledge and language development. Examinees are asked to identify pictures of familiar and unfamiliar objects and to respond to antonyms, synonyms, and verbal analogies.</td>
</tr>
<tr>
<td>2. Visual-Auditory Learning (Glr)</td>
<td>Measures visual-auditory association. Examinees must associate new visual symbols with familiar words and translate a series of symbols into verbal sentences.</td>
</tr>
<tr>
<td>3. Spatial Relations (Gv)</td>
<td>Measures visual processing. Examinees select from a visual list the component parts needed to complete the whole shape.</td>
</tr>
<tr>
<td>4. Sound Blending (Ga)</td>
<td>Measures auditory phonics coding. Examinees must integrate and say whole words after hearing parts of them.</td>
</tr>
<tr>
<td>5. Concept Formation (Gf)</td>
<td>Measures ability to deduce relations. Examinees identify the rules for concepts when shown illustrations of both instances and noninstances of the concepts.</td>
</tr>
<tr>
<td>6. Visual Matching (Gs)</td>
<td>Measures speed in visual discrimination. Examinees visually identify two numerals that are alike in a row of six numbers. As the items progress, the task becomes increasingly difficult.</td>
</tr>
<tr>
<td>8. Incomplete Words (Ga)</td>
<td>Measures auditory closure. Examinees must state a complete word after listening to a word missing one or more phonemes.</td>
</tr>
<tr>
<td>9. Auditory Working Memory (Gsm)</td>
<td>Measures working memory. Examinees are first asked to listen to a mixed set of numbers and words, then put this series in sequential order, words first, then numbers.</td>
</tr>
<tr>
<td>10. Visual-Auditory Learning-Delayed (Glr)</td>
<td>Measures delayed recollection after 1 to 8 days. Examinees are presented with symbols from Visual-Auditory Learning, state the name of each, and are retested later, without knowing the retest will occur.</td>
</tr>
<tr>
<td>11. General Information (Gc)</td>
<td>Measures general verbal information and contains two subtests. In the first subtest, examinees are asked where one would find a certain object. In the second subtest, examinees are asked what one would do with this object.</td>
</tr>
<tr>
<td>12. Retrieval Fluency (Glr)</td>
<td>Measures ideation fluency. Examinees are given a category and asked to name as many examples from that category as possible in 1 minute.</td>
</tr>
<tr>
<td>13. Picture Recognition (Gv)</td>
<td>Measures visual memory. Examinees must recognize a subset of previously presented pictures within a field of distracting pictures.</td>
</tr>
<tr>
<td>14. Auditory Attention (Ga)</td>
<td>Measures speech-sound discrimination and resistance to auditory stimulus distortion. Examinees are presented with four pictures and are asked to listen to a word that is audiotaped and superimposed over background noise. Examinees are then asked to point to the correct picture for that word.</td>
</tr>
<tr>
<td>16. Decision Speed (Gs)</td>
<td>Measures semantic processing speed. Examinees are presented with a row of pictures and are asked to identify the two pictures that are most similar.</td>
</tr>
<tr>
<td>17. Memory for Words (Gsm)</td>
<td>Measures serial auditory memory. Examinees repeat lists of unrelated words in correct sequences.</td>
</tr>
<tr>
<td>18. Rapid Picture Naming (Gs)</td>
<td>Measures naming facility and speed. Examinees are presented with a series of pictures and are asked to name them within a 2-minute time limit.</td>
</tr>
<tr>
<td>19. Planning (Gv)/(Gf)</td>
<td>Measures executive processing, spatial scanning, and general sequential reasoning. Examinees are asked to trace an outline without picking up their pencil or retracing lines.</td>
</tr>
<tr>
<td>20. Pair Cancellation (Gs)</td>
<td>Measures attention and concentration. Examinees are asked to identify and circle a repeated pattern in a 3-minute time limit.</td>
</tr>
</tbody>
</table>

Das-Naglieri Cognitive Assessment System (CAS)

The CAS (Naglieri & Das, 1996) is a measure of cognitive processing that uses a battery of tasks. It is very well researched, uses nontraditional approaches for assessing differences in intelligence (Meikamp, 2001), and, in general, reduces the emphasis on client verbal and language skills compared with other prominent intelligence tests. It has two formats, basic and standard. The Basic Battery has 8 subtests and can be administered in 40 minutes. The Standard Battery has 12 subtests and can be administered in 60 minutes. The CAS is individually administered and is appropriate for clients ages 5 to 17 years. The test is based on the PASS approach, developed by its authors, which is supported by the basic cognitive functions of Planning, Attention, and Simultaneous and Successive processing (Anastasi & Urbina, 1997). The Planning function includes the Matching Numbers, Planned Codes, and Planned Connections subtests. The Attention function includes the Expressive Attention, Number Detection, and Receptive Attention subtests. The Simultaneous Processing function includes the Nonverbal Matrices, Verbal-Spatial Relations, and Figure Memory subtests. The Successive Processing function includes the Word Series, Sentence Repetition, Speech Rates (ages 5 to 7 years), and Sentence Questions (ages 8 to 17 years) subtests. The inclusion of the Planning and Attention scales make this test unique compared with other traditional methods of intelligence assessment (Meikamp). Thompson (2001) reported that the directions for administration are straightforward, and it is easy to learn to administer. According to the test publisher (Riverside Publishing, 2005a), stratified standardization samples (n = 2,200) conformed to U.S. population demographics.

Raw scores for the CAS can be converted into scaled scores, standard scores ($M = 100; SD = 15$ for the Full Scale score), percentile ranks, and age equivalents. According to the Interpretive Handbook, score reliability is reportedly excellent, as the full-scale internal consistency coefficient was 0.96, and PASS scale reliabilities ranged between 0.83 and 0.93. The test-retest reliability coefficients ranged from $r = 0.64$ to $r = 0.92$, with an average interval of 21 days (Meikamp, 2001). Discriminant validity was supported in the Interpretive Manual, and other reliability and validity studies were reviewed (Riverside Publishing, 2005a). Meikamp warned that even though the CAS can help to gain a broader understanding of individual differences, comparisons of PASS scales and subtest scores may be easily misinterpreted, and further empirical research is necessary. Thompson (2001) concluded that he found the CAS to be a more superior measure of intelligence than the Stanford-Binet (Fourth Edition), but because of easier use and scoring interpretation, he preferred the Wechsler scales.

References:


