Chapter 2

Measurement Issues

Chapter Questions

- What type of information is yielded from a T-score?
- How does skewness affect scaling decisions?
- How has factor analysis been used to develop personality tests and diagnostic schedules?

Users of instruments assessing personality and other aspects of behavioral, emotional, and social functioning should have a thorough understanding of measurement principles. The discussion that follows, however, hardly qualifies as thorough because measurement instruction is not the purpose of this book. This chapter merely points out some of the most important measurement concepts for conducting assessments of youth.

We assume that the user of this text has had, at a minimum, undergraduate courses in statistics, tests and measurements, as well as at least one graduate-level measurement course. If a user of this text is not acquainted with some of the principles discussed here, then a statistics and/or measurement textbook should be consulted. There are a number of excellent measurement textbooks available, including Anastasi and Urbina (1998) as well as Allen and Yen (1979). The reader is also referred to the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) for a discussion of the appropriate procedures for test development, test selection, scoring, interpretation, and communication of results.

This chapter begins by defining the nature of the tests that assess psychological constructs. Then, a review of basic principles

P.J. Frick et al., *Clinical Assessment of Child and Adolescent Personality and Behavior*,
DOI 10.1007/978-1-4419-0641-0_2, © Springer Science+Business Media, LLC 2010

of statistics and measurement is presented, including topics ranging from measures of central tendency to factor analysis. The last part of the chapter introduces measurement issues that are specific to the use and interpretation of personality tests and similar instruments.

Defining Personality Tests

There is a plethora of methods, including tests, designed to assess similar-sounding psychological constructs, including personality scales, behavior rating scales, and diagnostic schedules. The available personality measures differ to such an extent that they can be subtyped in order to clarify their psychometric properties. A definition for a psychological test, taken from an early, well-known personality assessment text, may be a good starting point. Kleinmuntz (1967) defines a psychological (including personality) test by observing, "A psychological test is a standardized instrument or systematic procedure designed to obtain an objective measure of a sample of behavior" (pp. 27–28). This rather broad definition provides a useful starting point for conceptualizing the great variety of measures available.

The central characteristic of this definition is the notion of standardization of behavioral sampling. *Standardization* has at least two meanings: *standardization* in the sense of collecting a sample for the purpose of norm referencing and *standardization* as administration of the measure according to a consistent set of rules. Most of the measures discussed in this volume fit the first notion of standardization in that they are norm-referenced. That is, these measures use norm groups for gauging a child's performance in comparison to some reference group. Furthermore, the principle of administration structure or consistency applies to all of the measures in this text. For example, respondents should complete the measure in an environment free of distractions and should clearly comprehend the response-format (e.g., true/false, frequency ratings, etc.) and time frame (e.g., the last 6 months) referenced by the test. Standardized procedure emanates from experimental psychology, where laboratory control is central to obtaining reliable and valid results (Kamphaus, 2001). Similarly, in the case of personality assessment, standardized administration procedure is necessary to produce reliable and valid measurements of behavior.

All psychological tests take a sample of behavior from which the findings are subsequently generalized (Anastasi & Urbina, 1998). This ability to generalize findings is the central strength of psychological tests and is probably the reason for their widespread use. Without these tests, psychological measurement would be impractical because of the time and expense required. Of course, a sample can always be in error, a fact that should always be considered when interpreting results (Dahlstrom, 1993).

Types of Tests

How does one identify an instrument that assesses personality or behavioral, social, and emotional functioning? Personality tests have traditionally attempted to assess personality traits such as introversion, agreeableness, and anxiety. As noted in Chap. 1, traits are usually considered to be relatively stable characteristics of the individual (Martin, 1988). For children and adolescents, such characteristics may be similarly conceptualized under the term *personality traits* or, typically for younger children, *temperament*. Research has clearly indicated that individual differences in a variety of personality domains in youth are measurable (see Shiner & Caspi, 2003), and relatively stable (e.g., Durbin, Hayden, Klein, & Olino, 2007; Hampson, Andrews, Barckley, & Peterson, 2007).

Rating scales, one of the most popular child assessment methods, may fall into yet another category of test called diagnostic schedules (Kamphaus et al., 1995). Kamphaus et al. define a diagnostic schedule as a specialized psychometric method that provides a structured procedure for collecting and categorizing behavioral data that correspond to diagnostic categories or systems. A diagnostic schedule, then, is not designed to assess a trait, but rather to diagnose a syndrome. How does one identify a diagnostic schedule? One clue is the source of the item pool. The Children's Depression Inventory (CDI; Kovacs, 1992) is a popular measure of childhood depression that used the *DSM* as its item source. It was designed to assess the symptoms of depression in order to assist with making the diagnosis of depression. It was not designed to assess a stable personality trait or temperament but rather to allow the examiner to make the diagnosis of depression with confidence. In fact, a cut score that indicates the possible presence of clinical depression is often used for interpreting scores (Kovacs).

However, adding further complexity to understanding how rating scales fit within the array of tools available for clinical assessments is the fact that many widely used rating scales cannot be considered diagnostic schedules. For example, although the Behavior Assessment System for Children-2 (BASC-2; Reynolds & Kamphaus, 2004) assesses clinically relevant domains for youth (e.g., hyperactivity, aggression, depression, anxiety), elevated scores on those domains do not necessarily mean that the individual being assessed meets the criteria for a corresponding diagnosis. Such rating systems routinely have items that do not directly map onto the diagnostic criteria. Rather, the content of these scales may be indicative of aspects of the young person's functioning that may lend themselves to recommendations for intervention, as well as help signal a diagnosis or impairment in a particular domain. Furthermore, some rating scales may blend the elements of personality tests and diagnostic schedules, making sound clinical judgment essential in drawing the most sound conclusions from the data collected. The primary purpose of the assessment (e.g., diagnostic clarification vs. identifying areas of behavioral, social, or emotional concern) should guide the selection of diagnostic schedules and/or behavior rating scales. Further, we would argue that if given a choice, clinicians should initially seek tools that provide a broad screening of a variety of possible problems rather than narrowing in too quickly on a specific diagnosis.

Mash and Hunsley (2005) have articulated the problems with considering a focus on specific diagnoses as synonymous with psychological assessment:

> "Although formal diagnostic systems…provide one alternative for framing the range of disorders and problems to be considered, there is no need to limit the range of problems to those detailed in a diagnostic system. Refraining from excessive reliance on formal diagnostic systems is warranted given the well-documented shortcomings in the nature and development of such systems (e.g., Beutler & Malik, 2002; Mash & Dozois, 2003; Scotti, Morris, McNeil, & Hawkins, 1996) and the lack of evidence that such diagnostic systems provide the best way to match a treatment to a child (Bickman, 2002)" (p. 368).

Despite these concerns, diagnostic schedules or checklists may still play a critical role in helping to address a referral question and make treatment recommendations that are diagnostically-relevant (e.g., classroom accommodations for a child who meets criteria for ADHD). Diagnostic schedules have evolved from behavioral assessment methods, as has the *DSM*, which now emphasizes the tally of behaviors

(symptoms) in order to make a diagnostic decision. Personality tests and many rating scales, on the other hand, are rooted in the psychometric tradition in which such tests are designed to assess traits across a continuum. While such instruments may not lead directly to a diagnostic decision, as noted above, they can play other important roles by identifying traits that have implications for the course or prognosis of a disorder, or even for treatment.

While diagnostic schedules are practical for making diagnostic decisions, such measures have limitations for studying the nature of individual differences or for contributing to other important aspects of the assessment process. These limitations are inherent in diagnostic schedules because they often lack a clear theoretical basis or evidence of a priori defined trait(s) that can be supported with construct validity evidence. Therefore, the emergence of diagnostic schedules as the instruments of choice for much of assessment practice is evidence of the profound impact of behavioral-based diagnostic systems on psychometric test development, particularly over the last decade or two, as well as the (real or perceived) need to provide diagnoses as a result of all assessments due to managed health care.

Appropriate conclusions that could be drawn based on diagnostic schedules include statements like the following:

- Tonya suffers from major depression, single episode, severe.
- Tony exhibits nearly enough symptoms to be diagnosed as having conduct disorder.
- Traci has attention problems that are worse than those of 99% of the children her age.

Alternatively, conclusions that could be offered based on psychometric tests of personality or behavioral, emotional, or social functioning could include:

- Allison shows evidence of poor adaptability to new situations and changes in routine, which puts her at risk for school adjustment problems.
- Patrick's high score on the sensation seeking scale warrants consideration as part of his vocational counseling and educational planning.
- Maria's somatization tendencies reveal the need for counseling in order to reduce her frequency of emergency clinic visits.
- Andersen's apparent signs of depression indicate a need for further evaluation and intervention.

A central difference between these interpretive statements is that those made based on diagnostic schedules are dependent on diagnostic nosologies. A variation of this premise is the third statement exemplifying diagnostic-based conclusions, which may result from a norm-referenced behavior rating scale that has a scale devoted to inattention. Such norm-based information can typically be gleaned from personality tests or other rating scales as well. The interpretive statements made based on psychometric tests, however, can be offered independently of diagnosis. These conclusions are based on the measurement of traits or tendencies that may or may not represent diagnostic symptoms or signs, and yet, these conclusions contribute substantially to the assessment process.

Widely used rating scales such as those to be discussed later in this volume have several scales with the same name as a diagnostic category such as depression or anxiety. At the same time, such measures are scaled similarly to traditional personality tests with standard scores based on norms.

Although research is emerging on this issue (e.g., Ferdinand, 2008; Kerr, Lunkenheimer, & Olson, 2007), generally speaking, we do not know the extent to which these scales demonstrate the

stability associated with traits or the diagnostic accuracy of the *DSM* system. Their popularity for clinical practice, however, continues to increase due to their cost effectiveness and time efficiency. Furthermore, rating scales allow for the rapid and accurate identification of domains of deviant behavior that may require diagnosis or treatment (Hart & Lahey, 1999).

In this volume, the term *personality test* will occasionally be used generically to apply to personality trait measures, diagnostic schedules, syndrome scales, and related measures, always assuming that the reader is aware of the distinctions between subtypes of measures.

Scores, Norms, and Distributions

Types of Scores

In this section, some of the basic properties of score types are reviewed, with particular emphasis on the T-score standard score metric and its variants. The properties of these scores will be highlighted in order to encourage psychometrically appropriate score interpretation.

Raw Scores

The first score that the clinician encounters after summing item scores is usually called *a raw score*. Raw scores, on most tests, are simply the sum of the item scores. The term *raw* is probably fitting for these scores in that they give little information about a child's performance as compared to his or her peers. Raw scores are not particularly helpful for norm-referenced interpretation. Raw scores merely identify the number of behaviors or symptoms present, not how deviant this amount of symptomatology is from the norm nor how impairing it is for the individual.

Norm-Referenced Scores

Personality test interpretation often focuses on *norm-referenced interpretation*, the comparison of children's scores to some standard or norm. For the purposes of assessing psychological constructs, scores are usually compared to those of children the same age. Norm-referenced achievement tests, by contrast, may compare children's scores to those of others in the same grade, and college admission counselors may compare an incoming student's GPA to that of freshmen who entered the year before.

Norm referencing is of importance in personality and behavioral assessment because it allows the clinician to gauge deviance, which is often central to the referral question. Parents who refer a child for a psychological evaluation often have norm-referencing in mind. They ask questions such as "Is her activity level normal for her age?" or "Everyone says he is just a boy, but fire setting isn't normal, is it?" Norm-referencing allows the clinician to answer such questions objectively. The remaining scores discussed in this section are norm-referenced scores that allow the clinician to make these important comparisons.

Standard Scores

The *standard score* is a type of derived score that has traditionally been the most popular for psychometric test interpretation. Standard scores convert raw scores to a distribution with a set mean and standard deviation and with equal units along the scale (Anastasi & Urbina, 1998). The typical standard score scale used for personality tests and behavior rating scales is the T-score, which has a mean of 50 and standard deviation of 10. Another popular standard score that is coming into more frequent use for personality test interpretation has the mean set at 100 and the standard deviation at 15, similar to the IQ metric (see Table 2.1). Because they have equal units along the

TABLE 2.1 Standard Score, T-Score, Scaled Score, and Percentile Rank Conversion Table

Standard Score M = 100 SD = 15	T-Score M = 50 SD = 10	Scaled Score M = 10 SD = 3	Percentile Rank	Standard Score M = 100 SD = 15	T-Score M = 50 SD = 10	Scaled Score M = 10 SD = 3	Percentile Rank
160	90		99.99	128	69		97
159	89		99.99	127	68		97
158	89		99.99	126	67		96
157	88		99.99	125	67	15	95
156	87		99.99	124	66		95
155	87		99.99	123	65		94
154	86		99.99	122	65		92
153	85		99.98	121	64		92
152	85		99.97	120	63	14	91
151	84		99.96	119	63		89
150	83		99.95	118	62		88
149	83		99.94	117	61		87
148	82		99.93	116	61		86
147	81		99.91	115	60	13	84
146	81	19	99.89	114	59		83
145	80		99.87	113	59		81
144	79		99.84	112	58		79
143	79		99.80	111	57		77
142	78		99.75	110	57	12	75
141	77		99.70	109	56		73
140	77	18	99.64	108	55		71
139	76		99.57	108	55		69
138	75		99	107	55		67
137	75		99	106	54		65
136	74		99	105	53	11	65
135	73	17	99	104	53		62
134	73		99	103	52		57
133	72		99	102	51		55
132	71		98	101	51		52
131	71		98	100	50	10	50
130	70	16	98	99	49		48
129	69		97	98	49		45

(Continues)

TABLE 2.1 (Continued)

Standard Score M = 100 SD = 15	T-Score M = 50 SD = 10	Scaled Score M = 10 SD = 3	Percentile Rank	Standard Score M = 100 SD = 15	T-Score M = 50 SD = 10	Scaled Score M = 10 SD = 3	Percentile Rank
97	48		43	68	29		2
96	47		40	67	28		1
95	47	9	38	66	27		1
94	46		35	65	27	3	1
93	45		33	64	26		1
93	45		31	63	25		1
92	45		29	63	25		1
91	44		27	62	25		1
90	43	8	25	61	24		.49
89	43		23	60	23	2	.36
88	42		21	59	23		.30
87	41		19	58	22		.25
86	41		17	57	21		.20
85	40	7	16	56	21		.16
84	39		14	55	20	1	.13
83	39		13	54	19		.11
82	38		12	53	19		.09
81	37		11	52	18		.07
80	37	6	9	51	17		.06
79	36		8	50	17		.05
78	35		8	49	16		.04
78	35		7	48	15		.03
77	35		6	48	15		.02
76	34		5	47	15		.02
75	33	5	5	46	14		.01
74	33		4	45	13		.01
73	32		3	44	13		.01
72	31		3	43	12		.01
71	31		3	42	11		.01
70	30	4	2	41	11		.01
69	29		2	40	10		.01

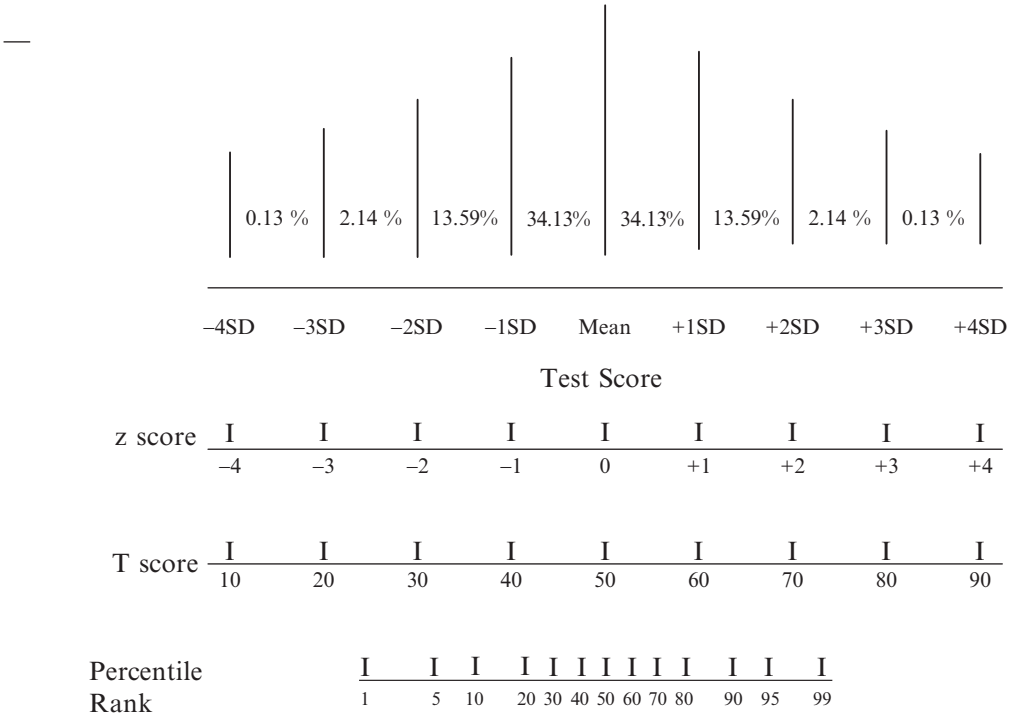

FIGURE 2.1
A normal distribution of scores

scale, standard scores are useful for statistical analyses and for making comparisons across tests. The equal units (or intervals) that are characteristic of standard scores are shown for various standard scores and percentile ranks in Table 2.1. This table may also be useful for converting a score from one scale to another. In the T-score metric, the distance between 20 and 30 is the same as that between 45 and 55.

Standard scores are particularly useful for test interpretation because they allow for comparisons among various subscales, scales, and composites yielded by the same test, allowing the clinician to compare traits. In other words, standard scores allow the clinician to answer questions such as "Is she more anxious than depressed?" thus facilitating profile analysis. Most modern personality tests use T-scores.

In a normal distribution (a frequently untenable assumption in personality and behavior assessment, as is shown in a later section), a normalized standard score divides up the same proportions of the normal curve. However, because many scales on syndrome measures in particular are heavily skewed (the most frequent scenario is that most individuals are not experiencing psychopathology and a few are, resulting in positive skewness), some test developers opt for the use of linear T-scores. Linear T scores maintain the skewed shape of the raw score distribution, which means that the same T-score on different scales may divide up different portions of the norming sample. Specifically, 50% of the norming sample may score below a linear T-score of 50 on the Anxiety scale, whereas 55% of the norming sample may score below a linear T-score of 50 on the

Aggression scale. Essentially, then, the use of linear T-scores makes the relationship of percentile ranks to T-scores unique for each scale.

Percentile Ranks

A percentile rank gives an individual's relative position within the norm group. Percentile ranks are very useful for communicating with parents, administrators, educators, and others who do not have an extensive background in scaling methods (Kamphaus, 2001). It is relatively easy for parents to understand that a child's percentile rank of 50 is higher than approximately 50% of the norm group and lower than approximately 50% of the norm group. This type of interpretation works well so long as the parent understands the difference between the percentile rank and the percent of items passed.

Figure 2.1 shows that percentile ranks have one major disadvantage in comparison to standard scores. Percentile ranks have unequal units along their scale. The distribution in Fig. 2.1 (and in Table 2.1) shows that the difference between the 1st and 5th percentile ranks is larger than the difference between the 40th and 50th percentile ranks. In other words, percentile ranks in the middle of the distribution tend to overemphasize differences between standard scores, whereas percentile ranks at the tails of the distribution tend to underemphasize differences in performance (Kamphaus, 2001).

Here is an example of how confusing this property of having unequal units can be. A clinician would typically describe a T-score of 55 as average. When placed on the percentile rank distribution, however, a T-score of 55 corresponds to a percentile rank of 69 in a normal distribution of scores (see Table 2.1). The percentile rank of 69 sounds as though it is higher than average. Examples such as this clearly show the caveats needed when dealing with an ordinal (unequal scale units) scale of measurement such as the percentile rank scale. It is important to remember that the ordinal properties of the scale are due to the fact that the percentile rank merely places a score in the distribution. In most distributions, the majority of the scores are in the middle of the distribution, causing small differences between standard scores in the middle to produce large differences in percentile ranks.

Uniform T-scores

A uniform T-score (UT) is a special type of T score that was used for development of the MMPI-2 norms (Tellegen & Ben-Porath, 1992). This derived score is a T-score like all other normalized standard scores with the exception that it maintains some (but not all) of the skewness of the original raw score distributions. The UT is like a normalized T-score in that the relationship between percentile ranks and T-scores is constant across scales, and it resembles a linear T-score metric in that some of the skewness in the raw score distribution is retained. The problem of a lack of percentile rank comparability across scales is described by Tellegen and Ben-Porath (1992) in reference to the MMPI-2:

> "For example, the raw score distribution of Scale 8, Schizophrenia (Sc), is more positively (i.e., right-) skewed than that of Scale 9, Hypomania (Ma). This means that a linear T-score of, say, 80 represents different relative standings on these two scales in the normative sample. For women in the MMPI-2 normative sample, the percentile values of a linear T-score of 80 are 98.6 for Scale 8 and 99.8 for Scale 9; for men, the corresponding values are similar, 98.6 and 99.7" (p. 145).

In order for the UT scale score to have the properties of percentile rank comparability across scales and reflection of raw score distribution skewness, the UT-score is based

on the average skewness value across all of the clinical scales (Tellegen & Ben-Porath, 1992). This approach meets the objectives outlined by the developers, but it relies on the assumption that the skewness of the MMPI-2 clinical scales is similar. There is, however, evidence that some MMPI-2 clinical scales (e.g., Hypochondriasis and Schizophrenia) are far more skewed than others (see Tellegen & Ben-Porath). It appears that the UT is a compromise metric that meets test development objectives while, at the same time, not addressing completely the issue of different skewness across scales. More research and clinical experience with the UT metric is necessary to determine whether or not this method should be adopted by other test developers.

Norm Development

Sampling

Norm development is one area in which the technology of personality and behavioral assessment has generally lagged behind that of intelligence or achievement testing (Martin, 1988). Intelligence tests, for example, have routinely collected stratified national samples of children to use as a normative base. Stratification is used to collect these samples in order to match, to the extent possible, the characteristics of the population at large. Common stratification variables include age, gender, race, geographic region, community size, and parental socio-economic status (SES; Kamphaus, 2001). These variables are used presumably because they are related to score differences. Of these widely used stratification variables, SES is known to produce the most substantial score differences on intelligence measures (Kamphaus). The precedent, then, is set for the norming of personality and behavioral assessment tools.

This precedent, however, has not been followed in several important respects. Until recently, many relatively popular personality scales have not done a good job of stratifying their samples. Some norming samples do not control for geographic region, and others fail to control for SES. The result is a normative standard of unknown utility. While poor norming is less likely to be tolerated in intelligence and academic achievement assessment, it is less frequently criticized or even noted in discussions of personality assessment. We will, however, note the characteristics of norming samples in subsequent sections of this text. This is important because users of personality and behavior tests should know the characteristics of a test's norming sample in order to make the best decisions and gauge the amount of confidence to place in the obtained scores.

Intelligence, achievement, and adaptive behavior tests typically feature interpretation based on a national norm sample. In contrast, a national normative standard has often not been offered for personality tests. A substantial number of personality tests offer only local norms, a subset of the national normative sample. Local norms answer different questions than do national norms. Hence, their potential utility has to be evaluated prior to test selection and interpretation.

Local Norms

Local norms, or norms based on a specific population in a specific setting or location, may sometimes be more useful than national norms, particularly in terms of their relevance for the clinician's work, and in some cases, recency relative to national norms (Elliott & Bretzing, 1980; Petersen, Kolen, & Hoover, 1989). In order for local norms to be meaningful, however, the range of their usefulness must be defined clearly.

Regardless of the use of local or national norms, typical norm-referenced questions of interest to psychologists are diagnostic ones. Common questions might include:

- Does Lindsey have attention-deficit/hyperactivity disorder?
- Is Jose clinically depressed?
- Is Stephanie more anxious than other children her age?

One of the goals of diagnostic practice is consistency, which is fostered by the publication of diagnostic criteria. Consistent methods of diagnosis allow clinicians to communicate clearly with one another. If, for example, Dr. Ob Session in Seattle says that a patient is suffering from conduct disorder, then Dr. Sid Ego in Atlanta will know what to expect from this adolescent when he enters his office for follow-up treatment.

National norms similarly promote consistency. If a clinician concludes that a child has clinically significant attention problems based on a deviant score on an inattention scale, then others may reasonably conclude that this child has attention difficulties that are unusual for her age. Popular tests, however, may offer different local norms that can hamper consistent communication. Similarly, local norms also will be less generalizable to the general population of children who do or do not meet criteria for a particular diagnosis. The clinician must then balance these disadvantages of local norms with the potential for local norms to be more relevant to the population with which he/she works.

Gender-Based Norms

Personality and behavior measures are unusual in that gender-referenced (local) norms are sometimes offered by test developers. This practice is unusual in comparison to other domains of assessment where, although significant gender differences exist, national combined gender norms are typically the only ones provided. Specifically, intelligence, academic achievement, and adaptive behavior scales produce mean score differences between gender groups, but local norms by gender are rarely offered. Why then are gender local norms commonly offered for personality tests? Tradition could be the most parsimonious explanation.

When comparing a child to his or her gender group, the effects of gender differences in behavior are removed. Another way of expressing this is to say that, when gender norms are utilized, roughly the same proportion of boys as girls is identified as having problems. Because, for example, boys tend to have more symptoms of hyperactivity than girls (*DSM-IV*, APA, 1994), the use of gender local norms would erase this difference in epidemiology. Gender norm-referencing would identify approximately the same percentage of girls and boys as hyperactive, such that a boy would require more severe symptomatology to be identified as elevated on hyperactivity relative to other boys. Depression is another example of how gender norms may affect diagnostic rates. Much evidence suggests that girls express more depressive symptomatology than boys in adolescence (Weiss & Weisz, 1988). The use of gender norms for a depression scale would result in the same number of adolescent boys as girls exceeding a particular cut score, whereas general national norms would retain the known greater prevalence among adolescent girls.

Are gender local norms a problem? Not so long as clinicians are clear about the questions they are asking. A gender norm question would be "Is Traci hyperactive when compared to other girls her age?" whereas a national norm question would be "Is Traci hyperactive in comparison to other children her age?" General national norms are preferred when a diagnostic question is asked. An example of a diagnostic question is, "Does Frank have enough symptoms of depression to warrant a diagnosis?" The *DSM-IV* diagnostic criteria do not have differing thresholds for boys, so a gender norm would be inappropriate.

Age-Based Norms

Because there are substantial differences across age groups, intelligence and academic achievement tests routinely offer norms separately by age groups, typically using age ranges of 1 year or less. By contrast, age ranges as large as 5–7 years are frequently used for personality tests. This tradition of articulating norms for larger age groups may be attributable to personality traits often having smaller normative samples than intelligence and achievement tests and a lack of age group differences in personality and behavior characteristics (Martin, 1988). Some data suggest that the latter explanation may be more appropriate. That is, differences between adjacent age groups are often insignificant for behavior rating scales, whereas more meaningful differences only occur over longer developmental periods (e.g., Reynolds & Kamphaus, 2004).

Clinical Norms

A more unique norm group is a sample of children who have been previously diagnosed as having a mental health problem. This clinical norm-referenced comparison can answer questions such as:

- How aggressive is Sheila in comparison to other children who are receiving psychological services?
- Are Tonya's psychotic symptoms unusual in comparison to other children who are referred for psychological evaluation?

There is not a clear precedent for the development of clinical norms. The relevant demographic stratification variables have not been identified, making it difficult to judge the quality of clinical norms. Should clinical norms, for example, attempt to mimic the epidemiology of childhood disorders including 10% depression cases, 5% ADHD cases, and so on? Should norms attempt to match the epidemiology of specific disorders within child clinic-referred populations (i.e., include mostly externalizing disorders)? Should norms be offered separately by diagnostic category to offer a more exact comparison? Or should attempts be made to address each of these issues?

Until such standards emerge, clinicians should seek clinical norms that are at least well-described. A clear description of the sample will allow the clinician to determine if the clinical norm group has the potential to answer questions of interest. For example, the clinician who works in an inpatient setting may have more interest in a clinical sample of inpatients, whereas others may prefer that clinical norms be based on a referral population. If the clinical norm group for a test is not well-described, the clinician cannot meaningfully interpret the norm-referenced comparisons.

The *normal curve* refers to the graphic depiction of a distribution of test scores that is symmetrical (normal), resembling a bell. In a normal distribution, there are a few people with very low scores (these people are represented by the tail of the curve on the left in Fig. 2.1), a few with very high scores (the tail on the right), and many individuals with scores near the average (the highest point in the curve).

When a distribution is normal or bell-shaped, as is the case in Fig. 2.1, the standard deviation always divides up the same proportion. Specifically, ±1 standard deviation always includes approximately 68% of the cases in a normal distribution, and ±2 standard deviations always include approximately 95% of the cases. The normal curve is also sometimes referred to as the normal probability, or Gaussian curve.

Normal distributions, however, cannot be assumed for personality tests or behavior ratings. While intelligence and academic tests often produce near-normal distributions, personality tests often produce skewed distributions. Examples

of skewed distributions are shown in Figs. 2.2 and 2.3. The distribution depicted in Fig. 2.2 is negatively skewed; a positive skew is shown in Fig. 2.3. A mnemonic for remembering the distinction between positive and negative skewness is to note that the valence of the skewness applies to the tail, when positive is on the right and negative is on the left.

It is understandable that diagnostic schedules and syndrome scales such as behavior rating scales produce skewed distributions. After all, only a small proportion of the population is experiencing a particular disorder at some point in time, and the majority of individuals are free of such symptomatology (positive skew). On the other hand, it is quite likely that the distributions for many adaptive skills or behaviors would be negatively skewed, in that the majority of the population would possess high levels of such skills, particularly with age (c.f., Sparrow, Cichetti, & Balla, 2005).

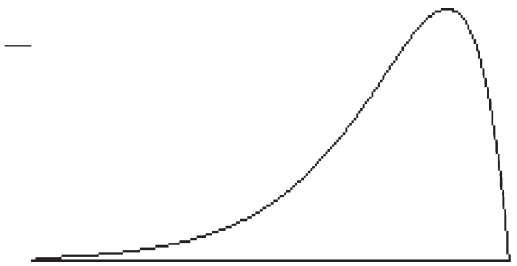

FIGURE 2.2
A hypothetical example of a negatively skewed distribution of scores

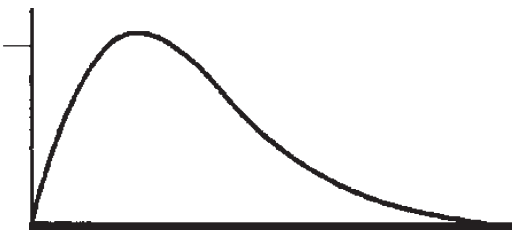

FIGURE 2.3
A hypothetical example of a positively skewed distribution of scores

The often skewed distributions obtained for personality measures, particularly diagnostic schedules, produce more controversy regarding scaling methods. If, for example, a distribution is heavily skewed, should normalized standard scores (which force normality on the shape of the standard score distribution regardless of the shape of the raw score distribution) or linear transformations (which maintain the shape of the raw score distribution) be used? Petersen et al. (1989) maintain that "usually there is no good theoretical reason for normalizing scores" (p. 226), and we concur with this opinion.

What differences does the scaling method make (i.e., normalized versus linear transformations)? The primary difference is in the relationship between the standard scores (T-scores) and percentile ranks yielded by a test. The positively skewed distribution shown in Fig. 2.3 is a good example of how this relationship can be affected. If this distribution was normalized (i.e., forced normal by converting raw scores to normal deviates and then the normal deviates to T-scores), then a T-score of 70 will *always* be at the 98th percentile. If linear transformations were used for the scale distribution shown in Fig. 2.3, then the corresponding percentile rank would most certainly be something other than 98. Clearly the type of standard score used for scaling a test affects diagnostic and, perhaps, treatment decisions. If normalized standard scores were used for a positively skewed scale (e.g., one measuring conduct problems), then potentially more children would be identified as having significant problems. On the other hand, normalized standard scores make the clinician's job easier by fostering interpretation across scales. Herein lies the debate: Is the inter-

pretive convenience of normalized standard scores worth the trade-off in lack of precision?

Clinicians will find that many tests use normalized standard scores (usually expressed in a T-score metric) even when clear evidence of significant skewness exists. We suggest that readers note the scaling method used by tests discussed in this volume as they consider the strengths and weaknesses of each measure.

Reliability

The *reliability* of a test refers to the degree to which test scores are free from measurement error and includes the presumed stability, consistency, and repeatability of scores for a given individual (AERA, APA, NCME, 1999).

The reliability of a personality test is expressed by the computation of a reliability coefficient, which is a special type of correlation coefficient. One essential difference between a reliability coefficient and a correlation coefficient is that reliability coefficients are typically not negative, while negative correlation coefficients are eminently possible. Reliability coefficients range, then, from 0 to +1. Reliability coefficients represent the amount of reliable variance associated with a test. In other words, a reliability coefficient is not squared, as is the case with correlation coefficients, to calculate the amount of reliable variance (Anastasi & Urbina, 1998). For example, the reliable variance of a test with a reliability coefficient of .90 is 90%, an unusually easy computation!

The error variance associated with a test is also easy to calculate. It is done by subtracting the reliability coefficient from 1 (perfect reliability). Taking the previous example, the error variance for a test with a reliability coefficient of .90 is 10% (1 - .90).

The sources of measurement error, while potentially crucial for interpretation, are often not specified, leaving the psychologist to engage in speculation. Error may result from changes in the patient's attitude toward assessment or cooperation, malingering, rater biases, patients' health status, subjective scoring algorithms, or item content that is incomprehensible to the examinee, among other factors.

For this reason, it is important to consider statistics that document both the reliable and error variance of a scale or test. In addition, multiple reliability coefficients and error estimates based on classical and modern test theory methods are necessary to guide clinical and research practice. Logically, then, it follows that no single estimate of reliability or error discussed in this section is adequate to support routine use of a test of assessment procedure.

Test–Retest Method

A popular method for computing the stability of personality test scores is the test-retest method. In this method the same test, for example the MMPI-A, is administered to the same group of individuals under the same or similar conditions over a brief period of time (typically 2–4 weeks). The correlation between the first and second administrations of the test is then computed, yielding a test-retest reliability coefficient that is optimally very close to 1.0. Of course, the importance of such reliability depends on the construct being assessed. If clinicians seek to assess changes in specific, discrete behaviors as a result of an intervention, for example, then test-retest reliability becomes less of a concern. On the other hand, if a clinician seeks to evaluate what are presumably relatively stable indicators of behavioral functioning or personality, the test-retest reliability of the measure becomes paramount.

Internal Consistency Coefficients

Another type of reliability coefficient typically reported in test manuals is an internal consistency coefficient. This estimate differs from test-retest or stability coefficients in that it does not directly assess the stability of the measure of personality over time. Internal consistency coefficients assess what the name implies-the average correlation among the items in a test or scale. In other words, this index of reliability assesses the homogeneity of the test item pool. Internal consistency coefficients are inexpensively produced, since they only require one administration of the test. Typical formula used for the computation of internal consistency coefficients include split-half coefficients, Kuder Richardson 20, and Coefficient (or Cronbach's) Alpha.

On occasion, there are differences between internal consistency and test-retest coefficients that can affect test interpretation. A test may, for example, have a relatively poor internal consistency coefficient and yet a strong test-retest coefficient (Kamphaus, 2001). Because internal consistency coefficients are imperfect estimates of stability coefficients, both types of coefficients should be recorded in the manual for a test (AERA, APA, NCME, 1999). It is then up to the professional making use of the test to determine if the reliability is suitable for the purpose for which the tool is to be used.

Variables that Affect Reliability

Clinicians who use personality or behavior tests should recognize factors that can affect reliability. Some factors that the clinician should keep in mind when estimating the reliability of a test for a particular child include the following:

1. Reliability can differ for different score levels. A test that is very reliable for emotionally disturbed students is not necessarily as reliable for nondisabled students without research evidence to support its use (AERA, APA, NCME, 1999).
2. Reliability can suffer when there is a long interval between assessments (Nitko, 1983).
3. Reliability can be affected by rater or child characteristics such as age, reading level, and fatigue. Reliability of personality measurement, for example, may drop if the child does not understand the test items.
4. Analogously, error may be introduced if a poor translation of a test is used.

Reliable Specific Variance

Subtest specificity is the amount of reliable specific variance that can be attributed to a single subtest or scale. Kaufman (1979) popularized the use of subtest specificity in clinical assessment as a way of gauging the amount of confidence a clinician should have in conclusions that are based on a single subtest. In effect, knowledge of subtest specificity makes clinicians more cautious about drawing conclusions based on a single scale.

A reliability coefficient represents the amount of reliable variance associated with a scale. An example would be an anxiety scale taken from a larger battery of 13 tests, all of which are part of a major personality test battery. The anxiety scale has a test-retest reliability coefficient of .82. On the surface, this test appears reliable. If this scale produces the child's highest score, the examiner may wish to say that the child has a problem with anxiety. The examiner can then make this statement with confidence because the test is relatively reliable, right? Not necessarily. As Kaufman (1979) points out, the conclusion being drawn by the clinician is about some skill, trait, or ability (in this case, anxiety) that

is specific or *measured only by this one scale.* The reliability coefficient, on the other hand, reflects not just reliable specific variance but also reliable shared variances. Subtest specificity is typically computed in the following way (Kamphaus, 2001):

1. Compute the multiple correlation (R) between the scale in question and all other scales in the battery, and square it (R^2). This computation yields the amount of reliable shared variance between the scale in question, in this case anxiety, and the other scales in the battery.
2. Subtract the squared multiple correlation coefficient from the reliability coefficient, or r_{tt}. If R^2= .30, .82 - .30 = .52. This formula yields the reliable specific variance.
3. Compare the amount of reliable specific variance (.52) to the amount of error variance (1 – .82 = .18). If the reliable specific variance exceeds the error variance by .20 or more, then the scale is considered to have adequate specificity for interpretive purposes. By convention, if the reliable specific variance exceeds the error variance by .19 or less, then the test lacks specificity, and it should be cautiously interpreted. If the reliable specific variance does not exceed the error variance, then interpretation of the scale is ill-advised.

Standard Error of Measurement

The standard error of measurement (SEM) gives an indication of the amount of error associated with test scores. In more technical terms, the SEM is the standard deviation of the error distribution of scores. The reliability coefficient of a test is one way of expressing the amount of error associated with a test score in order to allow the user to gauge the level of confidence that should be placed in the obtained scores. An examiner may report a personality test score for a child as being 63 with a test-retest reliability coefficient of .95. This practice, however, is unorthodox and clumsy. The typical practice is to report a test score along with the test's standard error of measurement, as is frequently done for opinion polls conducted by the popular media (e.g., the error rate or margin of error of this poll is…). The standard error of measurement is simply another way of reflecting the amount of error associated with a test score.

In classical test theory, if a child were administered a personality test 100 times under identical conditions, he or she would not obtain the same score on all 100 administrations. Rather, the child would obtain a distribution of scores that approximates a normal curve. This error distribution would have a mean. The mean of this theoretical distribution of scores is the child's true score. *A true score is a theoretical construct that can only be estimated.* This error distribution, like other distributions, not only has a mean, but it can also be divided into standard deviations. In an error distribution, however, instead of being called a standard deviation, it is called the SEM. As one would predict, then, in this error distribution of scores ±1 SEM divides up the same portion of the normal curve (68%) as does a standard deviation, and ±2 SEMs divide up the same proportion of the error distribution (95%) as ±2 standard deviations do for a normal distribution of obtained scores.

Confidence Bands

A confidence band is a probability statement about the likelihood that a particular range of scores includes a child's true score. As is done with opinion polls, clinicians use the SEM to show the amount of error, or unreliability, associated with obtained scores. Obtained scores are then banded with error. "Banding" is frequently accomplished by subtracting 1 SEM from, and adding 1 SEM to, the

obtained score. If, for example, the child obtained a T-score of 73 on the Reynolds Child Depression Scale (RCDS; Reynolds, 1989), one could apply the theory of standard error of measurement to band this score with error. For the total RCDS sample, the standard error of measurement rounds to 4 T-score points. Given that ±1 SEM includes approximately 68% of the error distribution of scores, the clinician could then say that there is a 68% likelihood that the child's true score lies somewhere in the range of 69–77. An examiner who wanted to use a more conservative ±2 SEMs could say that there is a 95% probability that the child's true score lies somewhere between 65 and 81. Confidence bands can be obtained for a variety of levels if one knows the SEM of the scale. Some manuals include confidence bands at the 68%, 85%, 90%, 95%, and 99% levels.

Construct Validity

Validity is defined as "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (AERA, APA, NCME, 1999, p. 184). There are a number of different ways of evaluating the validity of a test. Some of the more common types of validity evidence will be discussed in this section. Validity is the most important psychometric characteristic of a test. A test can be extremely well normed and extremely reliable and yet have no validity for the assessment of personality. One could, for example, develop a very good test of fine motor skill, but if one tried to make interpretations about someone's personality from this test, such interpretations would not be valid. That is, validity is essentially an issue pertaining to the uses of a test and the interpretations that one seeks to make from test results.

Virtually every aspect of a test either contributes to or detracts from its ability to measure the construct of personality or behavior, or, in other words, its *construct validity*. Construct validity is the degree to which a test measures some hypothetical construct. As such, the construct validity of a personality test cannot be established based on a single research investigation or the study of only one type of validity (e.g., factor analysis). Construct validity is based on the long-term accumulation of research evidence about a particular instrument, using a variety of procedures for the assessment of validity.

Based on the information provided in the previous paragraphs, it is clear that a statement that a test is valid or invalid is inappropriate. Instead, certain interpretations can have more or less evidence to support their validity and the accumulation of evidence in support of these interpretations is always ongoing (AERA, APA, NCME, 1999).

Content Validity

One of the reasons that many people would disagree with using a test of vocabulary knowledge as a measure of personality is that it does not appear to possess valid content. *Content validity* refers to the appropriate sampling of a particular content domain. Content validity has been most closely associated with the development of tests of academic achievement (Anastasi & Urbina, 1998). Typically, procedures for the establishment of content validity are judgmental (Petersen, Kolen, & Hoover, 1989).

Personality test developers have often relied on empirical test development methods, in which items are assigned to scales based on statistical properties only (such as factor loadings, to be discussed later), and many manuals do not provide a clear indication of the source of items. In

some cases, the item source is clear, such as with the Children's Depression Inventory (CDI; Kovacs, 1991), where items were based on accepted diagnostic nosologies such as the *DSM*. Even in such cases, however, personality test developers usually do not go to the lengths of other test developers to document adequate sampling of the content (or psychopathology) domain. Few personality tests or behavioral rating scales, for example, use panels of experts to develop item content.

Problems with regard to content validity may be identified as cases of construct underrepresentation or construct irrelevance. A depression scale may suffer construct underrepresentation, for example, if it lacks both cognitive (e.g., excessive self-deprecation) and vegetative symptoms of depression (e.g., problems sleeping). In this scenario, it may be said that there are not enough items on the scale that are known to be "indicators," or symptoms of depression, resulting in questionable content validity.

The reader will note in later chapters that construct irrelevant items are a more serious problem in behavior assessment. This problem is likely to occur when only empirical methods are used to construct scales and select items for scales (i.e., factor analysis). Examples of construct irrelevance are listed in Chap. 17 as they relate to the assessment of ADHD. In comparison to some others, ADHD is a well-studied condition with a widely agreed-upon set of symptoms (e.g., motor hyperactivity and inattention). What if, however, an item such as "My child is adopted" was placed on an inattention scale of a parent rating scale? As is noted in Chap. 17, such an item would likely be identified as a source of construct irrelevant variance for this scale, a source that would lead to a less valid assessment of attention problems for the child undergoing evaluation.

In our view, construct irrelevance and construct underrepresentation are likely to become problems at the item selection stage of test development. We, therefore, caution test users to carefully review the process of item selection and scale construction for each test that they utilize. By doing so, we think that clinicians will be better able to judge the implications of test content for interpretation.

Criterion-Related Validity

Criterion-related validity assesses the degree to which tests relate to other tests in a theoretically appropriate manner. There are two kinds of criterion-related validity: concurrent and predictive.

Concurrent Validity

This type of validity stipulates that a test should show substantial correlations with other measures to which it is theoretically related. One of the important criteria for the evaluation of personality or behavior measures since their inception has been that they show a substantial correlation with other indicators of psychopathology, such as well-validated tests or clinicians' ratings or diagnoses. The typical concurrent validity investigation involves administering a new behavior rating scale and an existing well-validated measure of psychopathology to a group of children. If a correlation of .20 is obtained, then the concurrent validity of the new test would be in question. A .75 correlation, on the other hand, would be supportive of the validity of the new test.

Predictive Validity

Predictive validity refers to the ability of a test to predict (as shown by its correlation) some later criterion. This type of research investigation is conducted very similarly to a concurrent validity study, with one important exception. The critical difference is that in a predictive validity study the new personality test is first

administered to a group of children, and then sometime in the future-perhaps two months, three months, or even six years-a criterion measure (such as clinicians' ratings of adjustment) is administered to the same group of children (see Verhulst et al., 1994).

Correlations with Other Tests

One can use correlations with other tests to evaluate the validity of a behavior or personality test. In a sense, this method is a special type of concurrent validity study. The difference is that the correlation is not between a personality measure and some criterion variable, such as clinicians' ratings of adjustment, but between a personality test and a measure of the same construct, another personality measure. For example, if a new test of anxiety is published, it should show a substantial relationship with previous measures, but not an extremely high relationship (Anastasi & Urbina, 1998). If a new personality test correlates .99 with a previous personality test, then it is not needed, as it is simply another form of an existing test and does not contribute to increasing our understanding of the construct of personality. If a new anxiety scale correlates only .15 with existing well-validated anxiety scales, it is also likely not to be a good measure of personality. New personality tests should show a moderate to strong relationship with existing tests, yet contribute something new to our understanding of the construct of interest.

Convergent/Discriminant Validity

Convergent validity is established when a scale correlates with constructs with which it is hypothesized to have a strong relationship. Discriminant validity is supported when a personality measure has a poor correlation with a construct with which it is hypothesized to be unrelated. These types of validity may be important to consider if there are no existing, well-normed, or relatively recent measures of a construct. That is, one may not be able to judge the criterion-related validity of a measure because no other suitable measures of that particular construct exist. Of course, convergent and discriminant validity are important indicators of validity for existing/established measures as well.

If one were assessing the convergent and discriminant validity of a measure of anxiety, one would expect high correlations with other measures of anxiety and moderate correlations with other measures of depression, given the well documented association between anxiety and depression (Klein et al., 2005). However, one would expect only minimal correlations between anxiety and measures of learning problems, thus providing support for its divergent validity.

Factor Analysis

Factor analysis is a popular technique for validating modern tests of personality that traces its roots to the work of the eminent statistician Karl Pearson (1901). Factor analysis has become increasingly popular as a technique for test validation. A wealth of factor-analytic studies dates to the 1960s when computers became available. Factor analysis is difficult to explain in only a few paragraphs. Those readers who are interested in learning factor analysis need a separate course on this technique and a great deal of independent reading and experience. A thorough discussion of factor-analytic techniques can be found in Gorsuch (1988). An introductory-level discussion can be found in Anastasi & Urbina (1998) and Kamphaus (2001).

Factor analysis is a data reduction technique that attempts to explain variance in the most efficient way. Most scales or items included in a test correlate with one another. It is theorized that this correlation is the result of one or more common factors. The purpose of factor analysis is to reduce the correlations between all scales

(or items) in a test to a smaller set of common factors. This smaller set of common factors will presumably be more interpretable than all of the scales in a personality test battery considered as individual entities.

Factor analysis begins with the computation of an intercorrelation matrix showing the correlations among all of the items or scales in a test battery. Most studies of behavior or personality tests use item intercorrelations as input. These intercorrelations then serve as the input to a factor-analytic program that is part of a popular statistical analysis package.

The output from a factor analysis that is frequently reported first in test validation research is a factor matrix showing the factor loading of each subtest on each factor. *A factor loading* is, in most cases of exploratory factor analysis, the correlation between a scale and a larger factor.[1] Factor loadings range from -1 to +1 just as correlation coefficients do. Selected factor loadings for the MMPI-A factor analysis of the standardization sample (Butcher et al., 1992) are shown in Table 2.2. A high positive correlation between a scale and a factor means the same thing as a high positive correlation between two scales in that they tend to covary to a great extent. One can see from Table 2.2 that the Hysteria scale is highly correlated with Factor 1, for example, and that Mania is not highly correlated with Factor 1, but it is highly correlated with Factor 2.

Once the factor matrix, as shown in Table 2.2, is obtained, the researcher must label the obtained factors. This labeling is not based on statistical procedures, but on the theoretical knowledge and perspective of the individual researcher. For the MMPI-A, there is general agreement as to the names of the factors. The first factor is typically referred to as general maladjustment and the second as overcontrol. The third and fourth factors are named after the scales with the highest loadings on each: social introversion and masculinity-femininity (Butcher et al., 1992).

Test developers often eliminate scales or items based on factor analyses. They also commonly design their composite scores based on factor-analytic results. This process was not followed in the development of the MMPI-A, as this test was developed long before the ready availability of factor-analytic procedures. Although the MMPI-A appears to be a four-factor test, it produces 10 clinical scale T-scores, and no composite scores corresponding to the four obtained factors are offered. More recently developed tests, such as the CBCL, made heavy use of factor analytic methods in the development of scale and composite scores (see Chap. 7).

Generally, consumers of factor-analytic research seek comparability between the factors and composite scores offered for interpretation. If there is, for example, a one-to-one relationship between the number of factors found and the number of composite scores produced, then the validity of the composite scores is likely enhanced.

Confirmatory Factor Analysis

The procedures discussed thus far are generally referred to as *exploratory factor-analytic procedures.* A newer factor-analytic technique is called *confirmatory factor analysis* (Kamphaus, 2001). These two factor-analytic procedures differ in some very important ways. In exploratory factor analysis, the number of factors to be yielded is typically dictated by the characteristics of the intercorrelation matrix. That is, the number of factors selected is based on the amount of variance that each factor explains

[1]When orthogonal (independent or uncorrelated) rotation techniques are used (and these techniques are very frequently used in test validation research), the factor loading represents the correlation between the subtest and a factor. This is not the case when oblique or correlated methods of factor analysis are used (Anastasi & Urbina, 1998).

TABLE 2.2 Selected MMPI-A Factor Loadings

	Factors			
	1 General Maladjustment	2 Social Overcontrol	3 Introversion	4 Masculinity Femininity
Hs	.77	.09	.31	.05
D	.69	–.23	.51	–.08
Hy	.88	–.15	–.22	–.15
Pd	.71	.28	.21	.19
Mf	.08	.01	.07	–.84
Pa	.70	.19	.23	.25
Pt	.52	.39	.67	.06
Sc	.61	.38	.53	.36
Ma	.31	.78	–.04	.33
Si	.19	.10	.91	.02

NOTE: These are varimax rotated factor loadings.
SOURCE: Adapted from Butcher et al., 1992.

in the correlation matrix. If a factor, for example, explains 70% of the variance in the correlation matrix, then it is typically included as a viable factor in further aspects of the factor analysis. If, on the other hand, the factor only accounts for 2% of the variance in a factor matrix, then it may not be included as a viable factor.

In confirmatory factor analysis, the number of factors is not dictated by data, but rather by the theory underlying the test under investigation. In confirmatory factor analysis, the number of factors is selected a priori, as well as the scales that load on each factor (Keith, 1990). The primary test in confirmatory factor analyses is the correspondence (i.e., fit) between the factor structure dictated a priori and the obtained data. If there is a great deal of correspondence between the hypothesized structure and the obtained factor structure, then the validity of the personality test is supported (hence the term *confirmatory*) and the theory is confirmed. If, for example, a researcher hypothesized the existence of four factors in a particular personality test, the confirmatory factor analysis will test how well the data from a specific sample conform to this hyopothesized test structure.

Thorough confirmatory factor-analytic studies use a variety of statistics to assess the fit of the hypothesized factor structure to the data. These statistics may include a chi-square statistic, goodness-of-fit index, adjusted goodness-of-fit index, or root mean square residual (RMR). Several statistics are desirable for checking the fit of a confirmatory factor analysis because all of these statistics have strengths and weaknesses. The chi-square statistic, for example, is highly influenced by sample size (Glutting & Kaplan, 1990).

Cluster Analysis

Similarly to factor analysis, cluster analysis attempts to reduce the complexity of a data set. In factor analysis, it is typical to try to reduce a large number of variables (e.g., items) to a smaller set. In cluster analysis, researchers are most often interested in grouping individuals (as opposed to variables)

into groups of people who share common characteristics. Ward's (1963) hierarchical agglomerative method is one example of a popular cluster-analytic technique.

Several steps are common to cluster-analytic techniques, including the following:

1. Collect a sample of individuals who have been administered one test yielding multiple scores or a battery of tests.
2. For each variable (e.g., depression scores), compute the distance between each pair of children.
3. These distances between each individual on each variable are then used to produce a proximity matrix. This matrix serves as the input for the cluster analysis in the same way that correlation or covariance matrices are used as input in factor analysis.
4. Apply a cluster-analytic method that sorts individuals based on the distances between individuals that were plotted in the proximity matrix. In simple terms, clustering methods in this step match individuals with the smallest distance between individuals on a particular variable.
5. This sorting process continues until groups of individuals are formed that are homogeneous (i.e., have profiles of scores of similar level and shape).
6. Just as in factor analysis, the researcher has to decide next on the number of clusters that is the most clinically meaningful. Statistical indexes are provided as an aid to the researcher in this step.

Cluster-analytic techniques are useful in psychopathology research for identifying subtypes of disorders or for designing diagnostic systems (Borgen & Barnett, 1987). Cluster-analytic techniques have frequently been applied to identify subgroups based on their performance on a particular personality measure (e.g., LaCombe et al., 1991).

In a series of investigations, Kamphaus and colleagues have used cluster analysis of large data sets to identify children with *subsyndromal* behavior problems (Huberty, Kamphaus, & DiStefano, 1997; Kamphaus, Huberty, DiStefano, & Petoskey, 1997; Kamphaus et al., 1999). These studies of elementary school children suggest that there are numerous children with profiles suggestive of functional impairment in school or at home who, nevertheless, are either not diagnosed or do not meet accepted diagnostic criteria. Thus, these cluster analyses helped to classify children without mental health diagnoses but who may require prevention or treatment.

Sensitivity and Specificity

Identification of a diagnosis is one of the primary reasons for conducting an evaluation. A test that is to be used for such a purpose should possess evidence of *sensitivity*, or the ability to identify true positives (i.e., the percentage of children who actually have the disorder). A prototypical study might involve administering an electronic measure of inattention to a group of children with ADHD and a group without any psychiatric diagnoses ("normals"). In this type of investigation, electronic measures of inattention often demonstrate good sensitivity by correctly identifying the vast majority of cases of ADHD, a finding that then triggers investigation of *specificity*. Specificity refers to the relative percentage of true negatives, or the correct identification of individuals who do not have the disorder as not having the disorder. This same measure of inattention may also identify only 50% of the nondiagnosed sample as "normal." Therefore, it may have demonstrated good sensitivity but inadequate specificity (i.e., a high rate of false positives). Electronic measures of inattention often produce results of this nature. In an exhaustive review of the

literature on such measures, Riccio and Reynolds (2003) have found that, while sensitivity is typically good, evidence of specificity is often poor.

In a later chapter, we will observe that the standards for this type of sensitivity and specificity have been raised considerably by the most recent *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999). Now tests must demonstrate the ability to differentiate among diagnostic categories-not just *between* a diagnostic group and normality. Unfortunately, few test manuals provide evidence of this nature and many journal articles test only the relatively easy distinction between some condition and normality. The clinician, however, routinely has the more difficult task of differentiating among diagnostic categories. Again, electronic measures of inattention have not shown good evidence of diagnostic group differentiation. In fact, Riccio and Reynolds (2003) concluded that when children with a number of problems are included, the proportion of children correctly classified drops significantly.

More recent work has focused on the development of an evidence-base that will improve problem specificity, or the ability to distinguish particular problems from each other (Mash & Hunsley, 2005). Clinicians routinely are faced with this task which is also often considered "differential diagnosis." The call for a larger more sound evidence base also raises awareness of *positive predictive power* (i.e., the ability of an item to correctly identify a child with a particular problem) and *negative predictive power* (i.e., the ability of an item to correctly identify a child without a problem; Pelham, Fabiano, & Massetti, 2005). More research on these issues can only serve to assist in clinical decision making, but increased evidence will still not replace clinical judgment in integrating information from a variety of sources that have varying degrees of validity.

Threats to Validity

Readability

An obvious, but easily overlooked, threat to validity is the lack of ability of the parent, teacher, or child to understand the personality test items. While concern is often expressed about the ability of children to read test items, parents may also have difficulty due to limited educational attainment or cultural or linguistic differences. Harrington and Follett (1984) found that most tests available at the time they conducted their study failed to address the issue in their test manuals. They provide several suggestions to the practitioner for screening informants in order to guard against readability serving as a threat to validity.

For parents, Harrington and Follett recommend having examiners read the test instructions for the informant and paraphrase. Children can be asked to read some items from the beginning, middle, and end of the instrument aloud so the examiner can gauge the child's reading skill.

Related to this point is the problem of *translational equivalence* or the degree to which a translation of a test is equivalent to its original language form (AERA, APA, NCME, 1999). Evidence of translational equivalence should be offered to reassure the test user that a threat to validity is not present.

Response Sets

A response set is a tendency to answer questions in a biased fashion, thus masking the true feelings of the informant. These response sets are often mentioned, and addressed in construction and interpretation, in some personality tests .

The *social desirability response set* is the tendency of the informant to respond to items in a socially acceptable way (Anastasi & Urbina, 1998). Some personality tests include items and scales to assess the potential effects of such a response set.

"I like everyone that I meet" might be an item on such scales. The *acquiescence response set* is the tendency to answer "true" or "yes" to a majority of the items (Kaplan & Saccuzzo, 1993). A third response set is called *deviation*, and it comes into play when an informant tends to give unusual or uncommon responses to items (Anastasi & Urbina, 1998).

Guarding Against Validity Threats

Personality and behavior tests often include other validity scales or indexes in order to allow the examiner to detect validity threats. Some tests include fake bad scales, which assess the tendency to exaggerate problems. Computer scoring of personality tests has allowed for the inclusion of consistency indexes. One such index allows the examiner to determine if the informant is answering questions in a predictable pattern. A consistency index might be formed by identifying pairs of test items that correlate highly. If an informant responds inconsistently to such highly correlated items, then his or her veracity may be suspect.

Examiners often also conduct informal validity checks. One quick check is to determine whether or not the informant responded to enough items to make the test result valid. Another elementary validity check involves scanning the form for patterned responding. A form that routinely alternates between true and false responses may reflect a patterning of responses.

One way to limit the influence of response sets is to ensure that informants are clear about the clinician's expectations. Some clients may also need to take the personality test under more controlled circumstances. If an examiner has reason to believe, for example, that a child is oppositional, then the self-report personality measure may best be completed in the presence of the examiner.

Utility

As described earlier, clinical utility is the "next frontier" in evidence-based assessment. By the time an assessment instrument is well-known and widely used in clinical settings, it usually has demonstrated adequate reliability and construct validity. However, as Mash and Hunsley (2005) describe, the question of utility or whether the instrument provides "psychologists with the kinds of information that can be used in ways that will make a meaningful difference in relation to diagnostic accuracy, case formulation considerations, and treatment outcomes" (p. 365) remains. This concept can also be applied to the inclusion of a particular informant in the assessment process.

In short, a rating scale, for example, may be a valid indicator of depression, but its utility indicates how valuable that particular rating scale is for an assessment of depression relative to other measures and relative to the cost (monetary and time) involved in administering it. Validity, including incremental validity (i.e., the improved assessment decision as a result of adding a measure), is a necessary condition for utility, and establishing such validity evidence for an assessment tool, and especially an entire assessment battery, is arduous. Nevertheless, various forms of validity evidence are likewise not sufficient for demonstrating utility. Ultimately, in addition to cost effectiveness, the clinician must take into account the assessment's role in translating to effective intervention and subsequent positive change for a child (Mash & Hunsley, 2005).

Calls to examine the clinical utility of assessment are not entirely recent (e.g., Hayes, Nelson, & Jarrett, 1987). However, give the current state of affairs, our discussion of utility is necessarily brief. As the move toward evidence-based assessment becomes strengthened by a larger collection

of empirical research and improved communication about assessment strategies, future volumes will hopefully be well poised to take on a detailed review of evidence on clinical utility.

Conclusions

Knowledge of psychometric principles is crucial for the proper interpretation of personality tests. As psychometrics become more complex, clinicians have to become increasingly sophisticated regarding psychometric theory. Because personality assessment technology has generally lagged behind other forms of child assessment, knowledge of psychometric theory must be considered more often by the clinician when interpreting scores.

Some personality tests, for example, do not include basic psychometric properties such as standard errors of measurement in the manual. Such oversights discourage the user from considering the error associated with scores, which is a basic consideration for scale interpretation. Omissions like this one are rare in academic and intelligence assessment. The application of the SEM is merely one example of the psychometric pitfalls to be overcome by the user of personality tests. This chapter ends, however, on an optimistic note. Newer tests and recent revisions are providing more evidence of validity and test limitations in their manuals.

Chapter Summary

1. A T-score is a standard score that has a mean of 50 and standard deviation of 10.
2. A percentile rank gives an individual's relative position within the norm group.
3. In order to select a representative sample of the national population for any country, test developers typically use what are called *stratification variables*.
4. *Local norms* are those based on some more circumscribed subset of a larger population.
5. The *reliability* of a test refers to the degree to which its scores are repeated over several measurements.
6. The *standard error of measurement* (SEM) is the standard deviation of the error distribution of scores.
7. *A confidence band is* a probability statement about the likelihood that a particular range of scores includes a child's true score.
8. The *reliability* of a test may differ for various score levels.
9. *Construct validity* is the degree to which tests measure what they purport to measure.
10. *Factor analysis* is a data reduction technique that attempts to explain the variance in a personality or behavior test parsimoniously.
11. In *cluster analysis* researchers are most often interested in grouping individuals (as opposed to variables) into clusters that share common behavior or traits.
12. Personality tests and behavioral rating scales often include other validity scales or indexes in order to allow the examiner to detect validity threats.
13. *Sensitivity* refers to the ability of a test to identify true positives and *specificity* to the ability of a test to identify true negatives.
14. *Clinical utility* concerns how well a particular tool provides necessary information and does so in a unique and cost-effective manner relative to other tools (or informants).