


**THE TESTING PROCESS**

Psychoeducational testing is a process in which skilled examiners carefully observe the actual performance of persons under standardized conditions.* The process incorporates elements of both science and art and draws on informal and formal assessment methods. From science we obtain rules that govern the collection, recording, and interpretation of data so as to

*As a matter of convenience, the terms children and child will be used in this chapter to refer to all age ranges covered by the WISC-IV.

Note: This chapter is adapted from “Assessment of Test Behaviors with the WISC-III,” which appeared in the previous edition of this book.
establish standardized methods that guide the use of a test in every setting in which it is applied. Alternatively, examiners' artistic qualities come from their extended experience in administering tests and working in other ways with children, youth, and adults. This artistic and/or clinical aspect of assessment is especially important when working with individuals whose dispositions make them difficult to test. No set of rules can govern the manner in which a test always is administered. As every clinician knows, children differ in their personal needs and test-related qualities. In addition, testing conditions vary from setting to setting.

Two examples are provided. Christine willingly comes with the examiner. They develop rapport easily. She is eager to help the examiner set up the testing room, listens attentively to directions, endeavors to do her best, and sustains a high level of interest and motivation throughout the examination. In contrast, David accompanies the examiner to the testing room only after considerable encouragement, seems distracted and inattentive, is uncooperative, and displays low levels of interest and motivation. The needs and test-related behaviors of Christine and David differ considerably. Keen observation skills are needed to assist in guiding the test's administration, in interpreting results from cognitive (i.e., intelligence and achievement) measures, and in deciding whether the test results are valid.

The two examples illustrate the importance of observing behaviors peripheral to scorables test responses. Examiners traditionally have relied on informal methods (e.g., observations, interviews) almost exclusively to better understand conditions that impact the test's administration and children's test behaviors. However, their use of formal observational methods, including instruments designed specifically for this purpose, is increasing.

This chapter reviews qualities that may impact a child's test performance and discusses in detail those qualities found through research to have a measurable influence on the performance. Discussion focuses on individually administered measures of cognitive abilities. Possible benefits in using test-behavior information are identified. Several measures currently used to record children's test performance also are discussed. The most widely used standardized measure of test performance, the Guide to the Assessment of Test Session Behavior (GATSB) (Glutting & Oakland, 1993), is described in some detail.

EXAMINER'S OBSERVATIONS ARE CRITICAL TO TEST USE

An examiner's observations are critical to all features of test use. Their observations enable them to accurately record children's responses to test items. Moreover, information obtained through observations enables examiners to better understand the manner in which children arrive at their answers and to identify cognitive and other personal strengths and weaknesses, thus facilitating test interpretations. Their observations also enable examiners to describe children's spontaneous behaviors while being tested, including their interpersonal and learning styles and other qualities that may directly or indirectly impact test performance. This information enables examiners to make needed modifications in the manner in which the test session is orchestrated, to construct systematic records of children's behaviors, to compare their observations with reports from others who know the child, to assist in interpreting the test results, and to draw comparisons between the child being tested and others who are similar in terms of salient qualities. Thus, the employment of observational methods is critical to the assessment process (Sattler, 1988, 2003).

SEVEN IMPORTANT QUALITIES THAT MAY IMPACT TEST PERFORMANCE

Professionals prepared to systematically observe behavior are keenly alert to various qualities important to the assessment process. Children's responses to a test's questions constitute the most central and important behaviors to which examiners should attend. For example, when administering the WISC-IV (Wechsler, 2003), examiners diligently observe and record a child's response to each test item.

In addition, examiners observe various other qualities that fall just beyond this central focus, ones that may facilitate or adversely impact children's test behaviors and thus their ability to demonstrate their best performance. The following seven qualities fall within this second important focus: conditions in the manner in which the test is administered, language qualities, personal readiness, motivation, and temperament. In addition, some test behaviors are associated with specific handicapping conditions. An examiner's knowledge of these qualities can greatly enhance the evaluation process.

Until recently, these seven qualities were assessed informally by the examiner who kept a watchful eye out for conditions that might impede test performance or jeopardize test validity. The somewhat recent development of standardized measures to assist in the assessment of test behaviors has aided the examiner, especially in the assessment of qualities related to rapport, personal readiness, motivation, temperament, and special conditions associated with some handicapping conditions. As will be noted, measures of test-taking behavior are intended to utilize and supplant and not substitute for well-honed observations skills.

TESTING ROOM CONDITIONS

Examiners are responsible for ensuring that the testing room provides a comfortable and distraction-free testing environment. Furniture should be of appropriate height and comfortable to the children. Their attention and concentration should not be attenuated by auditory and visual distractions. Young children, children with moderate to severe handicapping conditions, and those for whom testing is a new experience often need additional time to
become oriented to the testing room. Examiners must remain alert to signs that children are uncomfortable or distracted and, when present, work to alleviate problems. Examiners must ensure that physical conditions enhance the assessment process by creating conditions that are relatively standard from test to test, ones that enable the examinee to feel comfortable and relaxed, encourage suitable test-taking behaviors, and promote valid testing.

**LANGUAGE QUALITIES**

Language qualities also figure importantly. Most measures of cognitive abilities rely on language to form a bridge between the examiner and examinee, enabling them to communicate. Although the WISC-IV is not intended to assess language directly, the quality and nature of children’s language can facilitate or impede test use. Language qualities include both receptive (i.e., listening comprehension and reading) and expressive (oral expression and writing) skills and abilities together with pragmatic (i.e., functional) language features. Language also may reflect dialect differences, including the use of non-standard English. In addition, some children have little to no knowledge of English.

Examiners typically rely on informal observations and information from interviews together with prior test data when forming judgments of children’s language abilities. More formal assessment of language is warranted when language deficits and differences are apparent. In addition, modifications in test use may be needed when using them with children who exhibit deficits or differences in one or more of these language areas. Modifications will be needed when using tests with children not fluent in English. The examiner’s keen attention to children’s language-related test behaviors together with information provided by other sources enable them to conduct their work in a more effective and efficient fashion.

**PHYSICAL AND MOTOR QUALITIES**

Children’s physical and motor qualities also may impact test performance. Information on general health conditions (e.g., respiratory problems, cardiovascular conditions, chronic or acute diseases and illnesses) and muscle control should be acquired from a parent or other informed adult. Gross muscle control can impact the assessment of adaptive behaviors. Fine muscle control is important to writing and other physical manipulations important to finger dexterity. Control of muscles in the oral cavity is prerequisite to comprehensible speech. Examiners must remain alert to physical and motor qualities that may adversely impact children’s test performance.

**RAPPORT**

Rapport refers to the nature of the interpersonal relationships between the child and examiner. Good rapport is characterized by harmony, con-
Children who are strongly extroverted generally prefer to express their ideas verbally. Moreover, their ideas often become better known to them after they hear themselves express their thoughts. In contrast, those who are strongly introverted generally prefer to express their ideas after they have time to reflect on the question or in writing.

Children who express strong practical preferences often are very attentive to details and can memorize well. In contrast, those who express strong preferences for imaginative styles are less attentive to detail, more interested in theories, and are more inclined to have problems memorizing specific facts and stories.

Children who are inclined toward thinking preferences generally enjoy competitive activities and displaying their knowledge. In contrast, those who express strong preferences for feeling are most inclined to disdain competition and favor cooperation and to rely on personalized standards when evaluating others. Furthermore, emotions seemingly have an impact on self-regulation important to taking tests (Schultz & Davis, 2000).

Children who prefer organized styles like their lives to be well organized, problems resolved, and things settled as quickly as possible. They often appreciate the structure tests provide. In contrast, those with flexible styles are more inclined to postpone decisions, generally prefer fewer rules and regulations, and enjoy situations that are not highly organized. They may find the rules governing testing to be too confining.

Examiners often observe these and other temperament-related behaviors while testing children and youth. These qualities may impact children's test performance styles. Examiners are encouraged to administer a measure of temperament to children and youth (e.g., Oakland, Glutting, & Horton, 1996) so as to better understand their preferred styles and the impact these styles may have on their test and school performance.

MODERATE TO SEVERE HANDICAPPING CONDITIONS

Various alterations often are needed when testing children who evidence moderate to severe handicapping conditions (e.g., children with visual or auditory handicaps, mental retardation, cerebral palsy, autism, emotional difficulties). The nature of the needed alterations depends on the child's age, the nature and severity of the child's handicapping condition, and prior testing experiences. The use of testing accommodations for persons with disabilities is common (Pitoniak & Royer, 2001).

Popular textbooks on intelligence testing discuss these issues in considerable detail (Gregory, 1992; Kamphaus, 2001; Kaufman, 1994; Palmer, 1984; Sattler, 1988, 2001). Thus, with the exception of referencing research on test-taking behaviors of children with attention deficit disorders, test-taking behaviors associated with each handicapping condition are not discussed here. Nevertheless, examiners must remain attentive to the special needs of children with handicapping conditions and strive to make modifications in the testing process while maintaining standardized methods. Considerable ingenuity and experience may be required when testing children with severe and multiple disorders.

BACKGROUND INFORMATION AND PERSONAL QUALITIES

Some qualities identified previously (e.g., medical, motor, food, sleep, acuity, and language status) can be considered background variables (Bracken, 1991). Information on them should be acquired before tests are administered. This information may assist the examiner in planning for the evaluation, in preventing the occurrence of problems, and addressing them should they occur.

Other qualities consist of ethically relevant expressions of personal qualities that can be observed only during testing. These include rapport and children's readiness for testing, their personality, and motivation. Although scales designed to assess test-taking qualities may include information on background variables, personal qualities displayed during testing always should be the main focus.

DISTINGUISH BETWEEN COLLECTING AND EVALUATING TEST BEHAVIOR

Distinctions should be drawn between the process of collecting and evaluating test behaviors. When collecting this information, examiners focus on qualities thought to facilitate and impede the administration of standardized tests. Examiners using informal or formal methods to collect test-behavior data enjoy similar degrees of flexibility when selecting the behaviors on which they focus. However, those using formal measures to assess test-taking behaviors should include behaviors identified by research as being important. Those using informal methods are neither guided by this knowledge nor governed by this constraint.

THREE STANDARDS FOR EVALUATING INFORMATION

When evaluating information on children's test-session behaviors, examiners must decide the standard to use in judging whether the behaviors are suitable. Three standards may be used. Examiners may evaluate a child's test-taking behaviors in reference to their notions of perfection, potential, or from normative standards.
Perfection. Perfection refers to whether the child's test behaviors were impeccable and unblemished, conditions that rarely occur. In addition, few examiners would agree on the exact qualities that constitute a perfect administration or be able to judge them reliably.

Potential. Potential refers to whether the test behaviors were as good as can be expected, given the conditions found among the seven previously described qualities that may impact test performance. This standard also is difficult to form and thus to use knowledgeably and reliably.

Normative standards. Normative standards are derived from data acquired from nationally standardized, normed, and well-validated measures designed specifically to assess test-taking behaviors. The use of this standard does not preclude consideration of the other two.

Examiners who rely only on informal evaluation methods also must rely on standards of perfection and potential. Those who use formal (i.e., standardized and structured) evaluation methods have the added advantage of using normative standards when age-appropriate norms are available for the structured rating scale they are using.

All examiners do and should use observational methods to describe the seven previously identified test-related behaviors. Examiners differ in how they use this observational information in an evaluative format to form judgments about the suitability of a child's test-taking behaviors and the resulting validity of the test data.

SOME BENEFITS OF USING INFORMAL PROCESSES

Some benefits may occur from using informal processes to evaluate test behaviors. The primary advantage in using informal processes over standardized and structured measures to evaluate test-taking behavior may lie in their flexibility. They allow examiners to tailor their observations in light of each child's qualities. For example, the test-taking behaviors most likely to impact the performance of children who are visually impaired will differ somewhat from those who are autistic. However, this flexibility is available when both informal and formal evaluation procedures are used.

Some examiners have years of experience testing children with specific types of disabilities (e.g., autism, visual handicaps) and thus are able to judge whether a child's behaviors are similar to others who display the disability. Nationally standardized scales designed to assist in evaluating test-taking behaviors do not provide norms for the various handicapping conditions.

Moreover, examiners often resist change and tend to follow traditions. They often continue to use a battery of measures they were taught in graduate school. Standardized measures of test-taking behaviors are relatively new. Thus, many examiners were not introduced to them during their

graduate training. In addition, some examiners dislike the need to purchase test-taking scales as well as the structure they provide. Thus, their widespread use will require both time and knowledge of their benefits.

SOME DISADVANTAGES TO USING INFORMAL PROCESSES

Informal methods to evaluate test behaviors have a number of disadvantages. Eight are identified here.

QUALITIES OBSERVED MAY BE IRRELEVANT

Examiners differ in the test-taking qualities they believe are most important. Their clinical preparation on this topic often is very uneven. Some receive excellent coursework and supervision while others labor under inadequate instructional systems that reflect diminished resources for expensive clinical graduate programs. As can be expected, examiners differ in their knowledge as to what test behaviors are most important to record and how to evaluate this information. Items on standardized measures help overcome some of these differences in preparation by enabling examiners to focus on important test-taking behaviors.

QUALITIES ARE UNSUPPORTED BY RESEARCH

Examiner's information about test behaviors rarely is based on solid research. Information on informal methods generally is embedded within extensive discussions of test-taking behaviors (cf. Bracken, 1991; Culbertson & Willis, 1993; Gregory, 1992; Kamphaus & Reynolds, 1987). Despite the volumes written on this topic (Culbertson & Willis, 1993; Epps, 1988; Gregory, 1992; Jensen, 1980; Kamphaus & Reynolds, 1987; Kaufman, 1990; Kaufman, 1994; Palmer, 1983; Reynolds & Kamphaus, 1990a, 1990b; Salvia & Ysseldyke, 1988; Sattler, 1988, 2001; Simeonsson, 1996), the amount of research in well-respected publications on test-taking behaviors is meager and can be carried easily by a 3-year-old child. Thus, our scientific knowledge as to the qualities that constitute test-taking behaviors is inadequate. In addition, reliance on informal methods to evaluate test behaviors has contributed to this deficit. The availability of standardized measures of test-taking abilities is likely to lead to more research on this important topic and thus improved literature on this important component of assessment.

QUALITIES EMANATE FROM FOLKLORE

Related to the first two points, informal methods are difficult to replicate and often breed folklore (i.e., opinions that over time take the form of
widely held established fact). The use of informal methods to clinically assess test-taking skills has contributed to folklore about various test-related issues. For example, examiners often believe that behaviors people evidence while taking a test (e.g., shyness) express personal traits they are likely to display in their everyday life. As we will see later, there is little evidence for this widely held belief. The continued overreliance on informal methods prevents the validation of this and other clinical folklore.

**OBSERVATIONS ARE UNSTRUCTURED**

Informal methods lack standardized methodology to record and score important test-taking behaviors. Methods to record and evaluate test behaviors differ from examiner to examiner as well as within an examiner. These conditions jeopardize the reliable and valid collection of information and thus attenuate their use.

**EXAMINERS ARE LESS CREDIBLE**

Examiners who use informal measures are less able to justify their conclusions. They increasingly are being required to justify their findings to colleagues, while testifying, and in other legal and professional settings. They are likely to be asked to justify the validity of their test results and often face challenging questions as to the nature of the test conditions and the examinee’s behaviors. Failure to record these qualities at the time of test administration jeopardizes their ability to successfully face cross-examination. Reliance on informal methods also may further jeopardize their testimony.

**AGE-RELATED DIFFERENCES MAY BE OVERLOOKED**

Examiners may be insensitive to important age-related differences. Children display different test behaviors at different ages (Glutting & Oakland, 1993). In general, test behaviors improve with age. Examiners who rely on informal methods to evaluate test-taking behaviors may be unaware of subtle but important age-related differences.

**NORMS ARE LACKING**

Clearly, the greatest limitation in using informal measures lies in their lack of a normative basis for comparisons. The availability of properly stratified norms is consistent with commonly accepted standards (i.e., Standards for Educational and Psychological Testing, American Educational Research Association et al., 1999) for test use as well as the expectations of those who receive clinical services. Thus, attempts to evaluate test behaviors without the use of norms invite error and should be avoided when possible.

**FAILURE TO CO-NORM OBSERVATION SYSTEM WITH STANDARDIZED TESTS**

The co-norming of two or more tests is becoming more common. This process enables examiners to better utilize information from the tests through their knowledge of relationships between them. The process also enables testing companies to economize when standardizing tests. One measure of test-taking ability, the Guide to the Assessment of Test Session Behavior (GATS; Glutting & Oakland, 1993) has been co-normed with both the Wexler Intelligence Scale for Children III (Wechsler, 1991) and the Wexler Individual Achievement Test (WIAT; Wechsler, 1992), thus providing direct normative links with two widely used measures. Before discussing the GATS, other formal measures designed to assess test-taking qualities are identified.

**OTHER FORMAL MEASURES OF TEST-TAKING BEHAVIOR**

Many older clinicians were introduced to the importance of test behaviors by Caldwell (1951), who provided an integrated view of test session behavior. Caldwell also introduced the Test Behavior Observation Guide (Caldwell, 1951), one of the first formal (i.e., structured) measures of test-related qualities. It contains 19 items that identify preexisting background characteristics as well as 15 items that focus on observable test behaviors.

A number of standardized measures of intelligence include a structured test-behavior scale. Some examples are provided here.

The Stanford-Binet Observation Schedule (Terman & Merrill, 1960; Thorndike, Hagen, & Sattler, 1986) appeared as part of the test's record booklet beginning with the 1960 edition and continuing through to its most recent version. Rational analysis was used in selecting the items that assess five domains: attention, reaction during test performance, emotional independence, problem-solving behavior, and independence of examiner support. Despite its availability and widespread visibility, evidence as to the scale's reliability and validity may be found in only one study (Glutting & McDermott, 1988).

The Stanford Binet Intelligence Scale Fifth Edition (Roid, 2003) protocol provides for the recording of eight test behaviors (e.g., adequacy of English usage, motor abilities, vision, health, and general testing conditions).

The WISC-IV protocol provides for the recording of seven test behaviors: language, physical appearance, acuity problems, unusual behaviors,
attention and concentration, attitudes toward testing, and affect/mood. The first four focus on more general qualities while the last three focus on those identified through research on the GATS being as important.

The Leiter International Performance Scale-Revised (Roid & Miller, 2000) provides one of the more complete scales for recoding test-related behaviors. Its 49 items assess the following eight qualities: attention, organization/impulse control, activity level, sociability, energy and feelings, self-regulation, anxiety, and sensory reactivity. The items were selected, in part, to assess qualities displayed by children who exhibit some common psychological and social problems.

The Kaufman Integrated Interpretive System Checklist for Behaviors Observed During Administration of WISC-III Subtests (Kaufman, Kaufman, Dougherty, & Tuttle, 1994) was developed to measure behaviors specific to individual subtests from the WISC-III. Example items include these: "Has difficulty understanding the long verbal directions to nonverbal tasks such as Picture Arrangement and Coding" and "Was distracted when trying to repeat digits forwards and backwards or solve oral arithmetic items."

The Woodcock-Johnson Tests of Cognitive Abilities (Woodcock, McGrew, & Mather, 2001) protocol also provides for the recording of seven test behaviors: conversational proficiency, cooperation, activity, attention and concentration, self-confidence, care in responding, and responses to difficult tasks (Schrank & Read, 2002).

Additional methods that assess test-taking behaviors are not a part of existing tests and thus may be used with most measures of cognitive abilities. The Behavior and Attitude Checklist (Sattler, 2001) contains 41 items covering 12 domains. The domains with the largest number of items include work habits, attitudes toward test situation, and expressive language. The Test Behavior Checklist (Aylward & MacGruder, 1986) contains 18 items, 3 of which focus on preexisting conditions and 15 on test-related behaviors.

DEFICIENCIES IN THESE FORMAL SCALES

The structured nature of each of the foregoing scales makes them somewhat useful in codifying children’s test behaviors. Nevertheless, a number of qualities limit their utility. Threats to the interpretability of data from these and other tests of test-taking occur as a result of construct irrelevant variance. For example, they include domains and behaviors thought to be important to the assessment of test-taking behaviors. However, empirical support for including them is lacking. Additionally, these tests generally display construct underrepresentation. That is, domains relevant to the assessment of test-taking behaviors may not be included on the test. Thus, the tests cited previously generally do not meet the two standards of construct relevance and construct representation.

Sound observations consider relevant and verifiable aspects of child functioning, including normal development (Glutting, 1986; McDermott, Watkins, 1985). Most items on these structured test-behavior scales overlook normal adjustment (Glutting & Oakland, 1993). Instead, they largely limit to evaluating pathological symptoms and negative behaviors.

A potentially more serious problem relates to the identification of integral dimensions (i.e., scales) underlying item sets. The majority of structured, test-behavior instruments are composed of undifferentiated lists of symptoms, or rationically derived symptom “domains.” These test-behavior measures do not present empirical evidence in support of either a single unifying construct or for their separate domains.

Perhaps their greatest deficiency is the absence of norms, information needed when evaluating how one child’s behaviors compare to those of others. As a result, examiners are left to their own resources in determining when a given child’s test behavior is normal or exceptional.

EMPirical research on children’s test Behaviors

The validity of test observations rarely has been studied. Thus, until recently, clinicians had to rely on professional experience and wise judgments when evaluating test-taking behaviors. Conditions changed during the 1980s as researchers began to initiate a series of investigations designed to better understand the construct and criterion-related validity of test observations.

CONSTRUCT VALIDITY

A synthesis of research on test-behavior studies (Glutting, Youngstrom, Oakland, & Watkins, 1996) identified only two studies that examined the construct validity of children’s test behaviors (Glutting & McDermott, 1988; Glutting, Oakland, & McDermott, 1989). Each study reported the results of factor-analyzed data from existing formal scales of children’s test behavior to determine the psychometric properties of these instruments and simultaneously to identify the number and nature of integral domains (i.e., scales) underlying children’s test behaviors.

Our initial investigation (Glutting & McDermott, 1988) examined the Stanford-Binet Observation Schedule (Terman & Merrill, 1960; Thorndike et al., 1986). A factor analysis of the Stanford-Binet Observation Schedule yielded two dimensions that accounted for 54% of the variance. The larger of the two factors was identified as Avoidance and the smaller domain was identified as Inattentiveness.
The factor structure and reliability of Test Behavior Observation Guide-related test observation data obtained from children between the ages of 7 and 14 years (Glutting et al., 1989) yielded a three-factor model that explained 58% of the total item variance. Respectively, the names and ordering of the factors were as follows: Inattentiveness, Avoidance, and Uncooperative Mood. Internal consistency estimates, based on Cronbach's (1951) formula for coefficient alpha, were .88, .77, and .72, respectively.

Results across these two factor-analytic studies showed that empirically derived domains of children's test behaviors possessed only modest relationships with rationally derived domains (Glutting & McDermott, 1988; Glutting et al., 1989). More important, empirically derived domains were fewer in number and had demonstrated reliability. Thus, the findings served to demonstrate that structured symptom lists and scales developed through the rational analysis of children's test behaviors are likely to be unproductive for most referrals that examiners encounter and generally should not be used for these purposes.

CRITERION-RELATED VALIDITY

Criterion-related validity, when applied to test observations, focuses on relationships between children's test-session behavior and their scores on formal tests (e.g., cognitive abilities) or behaviors observed in other contexts. Intrasession validity, a term we coined, refers to the strength of association between measures of test-session behaviors and test scores from measures of cognitive abilities (cf. Glutting & McDermott, 1988; Glutting et al., 1989). Thus, intrasession validity examines the potential impact of test behaviors on children's formal scores on the WISC-III and WIAT as well as other individually administered measures of cognitive abilities and indicates the extent to which scores from these measures can be considered to be accurate.

The construct of ecological validity draws attention to the importance of examining the generalizability of test behaviors to diverse settings (cf. Neisser, 1991). We developed the term exosession validity, similar in meaning to the constructs of external validity or generalizability, to describe the degree of accuracy of using children's test-taking behaviors to predict their behavior in other situations (Glutting et al., 1989).

Our synthesis of previous test-behavior research (Glutting et al., 1996) found six studies on the topic of intrasession validity. The sources yielded a total of 33 correlation coefficients. An averaged coefficient was calculated according to the meta-analysis procedures recommended by Hunter, Schmidt, and Jackson (1982) and Rosenthal (1991). The overall relationship was \( r = -.34 \) between children's test behaviors and IQs obtained during the same test session. Thus, test-taking behaviors can impact test performance.

Four studies (Glutting et al., 1996) discussed exosession validity, producing 26 correlations. The average correlation was .18 between children's test behaviors and their conduct in other contexts (e.g., their classroom or community). Thus, these results indicate test-taking behaviors have little relationship with children's test behaviors in their classrooms and communities.

The pattern shows modest but meaningful levels of intrasession validity (average \( r = -.34 \)). Moreover, the magnitude of intrasession validity is higher than that found for exosession validity (average \( r = .18 \)). The findings of high intrasession validity provide important evidence that clinicians are able to utilize test observations to form judgments of children's test behaviors that are both reliable and valid. These findings also provide a foundation for establishing a formal measure of test-taking behaviors that, after norming, would enable examiners to more accurately acquire and interpret test observations. When used jointly with measures of cognitive abilities, information from a measure of test-taking behaviors could be used to help validate scores obtained from measures of cognitive abilities and to form inferences regarding children's test-taking qualities.

The evidence of limited exosession validity was not surprising. Behaviors that occur in natural settings are best understood by acquiring information from multiple sources and using multiple assessment methods that assess multiple traits displayed over time. Examiners should not expect to be able to describe complex peripheral behaviors only from observing a person in controlled testing situations. Evidence of limited exosession validity does not diminish the value of test observations. Instead, clinicians are encouraged to refrain from drawing conclusions as to the generalizability of test observations to conditions outside the immediate test situation. This inference is consistent with other information regarding the situational specificity of children's behavior. A meta-analysis of behavioral data (Achenbach, McConaughty, & Howell, 1987) demonstrated that much of the behavior observed by parents at home and teachers in school is contextually dependent and specific to the situation in which it occurs.

TEST BEHAVIOR OBSERVATION FORM: AN EMPIRICALLY DERIVED AND NORMED SCALE

The Test Behavior Observation Form (TOF) (McConaughty & Achenbach, 2004) is designed to assess behavioral and emotional problems in children, ages 2 to 18 years, based on their test-session behaviors. Behaviors assessed include withdrawn/depression, language/thought problems, anxiety, oppositional disorders, as well as inattentiveness and hyperactivity. The authors suggest the TOF data can be compared with profiles derived from the authors' other behavioral measures (e.g., Child Behavior Checklist).
THE GUIDE TO THE ASSESSMENT OF TEST SESSION BEHAVIOR FOR THE WISC-III AND WIAT

The Guide to the Assessment of Test Session Behavior for the WISC-III and WIAT (GATS B) (Glutting & Oakland, 1993) was constructed to overcome the shortcomings of other measures of children's test behaviors. The GATS B is a structured 29-item behavior-rating instrument designed to evaluate the test-session behavior of children quickly and reliably when administering measures of cognitive abilities. The instrument is brief and requires less than 5 minutes to complete, including scoring. Examiners rate GATS B items immediately after testing. As a result, the process of rating does not interfere with the recording of children's performance on the standardized test that is being administered, and the behavioral data are recorded while still easily recalled.

Children's behaviors are rated using a three-point scale (i.e., 2, 1, or 0) in reference to usually applies, somewhat applies, and doesn't apply. Higher raw scores denote inappropriate behavior. Raw scores are summed and converted to standard T scores (M = 50, SD = 10) according to three factor-based scales (Avoidance, Uncooperative Mood, Inattentiveness). In addition, a Total Score is obtained. The Total Score is a combination of scores from the GATS B's three other scales and also is expressed as a standard T score.

ITEMS

Example items are presented from the three factor-based scales. The examples make clear that the GATS B does not overlook normal adjustment; its items depict both appropriate and inappropriate behavior.

Avoidance
- "Shows marked interest in test activities"
- "Hesitates when giving answers"

Uncooperative Mood
- "Performance deteriorates toward end of testing"
- "Asks how soon testing will finish"

Inattentiveness
- "Listens attentively to directions and test items"
- "Attempts to answer before questions are completed"

NORMS

The GATS B was designed for use with children ages 6 years 0 months through 16 years 11 months. It was co-normed with both the WISC-III and the WIAT. Thus, unlike all previous measures for evaluating children's test behaviors, the GATS B alone was co-normed with tests of intelligence and achievement. As previously noted, norms are essential in determining when children's test behaviors are sufficiently aberrant, compared to others of similar age and experience, to affect the validity of test scores.

Each of the GATS B's two standardization samples (one each for the WISC-III and for the WIAT) included 640 children. The samples were stratified on the basis of the 1988 U.S. census data according to age, race-ethnicity, gender, and parent education. Furthermore, the samples were selected to ensure that children's overall intellectual abilities (M WISC-III FSIQ = 100, SD = 15) and their achievement levels (M WIAT Total Battery Composite = 100, SD = 15) matched those of the general population. Thus, ratings on the GATS B are typical of distributional means, standard deviations, kurtosis, and skewness found in measures of children's intellectual ability and achievement.

AGE-BASED STANDARD SCORES

An analysis of variance was conducted using the factor-based raw scores from the two standardization samples to determine whether norms could be collapsed over years or should be provided for each year for which the GATS B was designed. Results showed some age differences in test-related qualities, warranting the need for separate norms for three age groups: 6 to 8 years, 9 to 12 years, and 13 through 16 years. Consequently, norms for the GATS B respect age-related differences that occur in children's test behaviors.

RELIABILITY

Internal consistency reliability estimates for the GATS B are high. Alpha coefficients were calculated separately for the three age groups. Results showed an averaged coefficient of .92 for the Total Score and coefficients between .84 and .88 for the three factor-based scales. Stability estimates (M interval = 1 day) also are high. Results across the three age levels show an averaged .87 for the Total Score and .71 to .77 for the three factor-based scales.

CONSTRUCT VALIDITY

The GATS B was not developed using rational analysis. Instead, its scales were assembled according to the substantive or construct approach to test development (cf. Cronbach & Meehl, 1955). Thus, both theoretical and empirical issues were considered. Factor analytic findings from our earlier studies indicated that test behaviors do not form a single unitary construct and instead are governed by as many as three underlying dimensions: avoidance, uncooperative mood, and inattentiveness (Glutting & McDermott, 1988; Glutting et al., 1989). These findings served as the theoretical underpinning of the GATS B.
Both items and factor analyses were used with the WISC-III/GATSB standardization data (N = 640) to develop a final scale composed of 29 items. Initial items in the pool (n = 102) were deleted when they showed no appreciable loading on the three hypothesized factors and when an item loaded appreciably on a factor contrary to theory. Thus, retained items were required to show structural relationships (i.e., factor loadings) paralleling theoretical relationships postulated for children's test behaviors. In turn, this methodology increased the probability that GATSB items would contribute to important underlying behavioral constructs evident during testing.

Principal components analysis and principle axis factor analysis (using both orthogonal and oblique rotations) yielded three dimensions for the standardization sample: Avoidance, Uncooperative Mood, and Inattentiveness. These dimensions are theoretically congruent, align with findings from previous studies of children's test behaviors, and are similar to established dimensions for evaluating children's adjustment and well-being in other contexts (cf. Achenbach & Edelbrock, 1983; Quay, 1986).

THREE SECONDARY FACTORS

Avoidance

The first pattern found in the GATSB is directly related to task aversion and fearfulness. Therefore, the term Avoidance was selected to refer to these qualities. Strong conceptual links exist between this factor and a major constellation of behaviors children display both in home and school environments. The factor is similar to one labeled Anxiety Withdrawal by Quay (1986) and Internalizing by Achenbach and Edelbrock (1983). Eleven items load on the Avoidance domain. This factor captures the lion's share of the total variance (49%) and suggests that children's task engagement and/or avoidance are likely to have the largest impact on their obtained intelligence and achievement test scores.

Uncooperative Mood

The second GATSB factor, Uncooperative Mood, consists of eight items. It accounts for 11% of the total item variance on the GATSB and reflects children's improper initial adjustment, lack of cooperation, and/or need for praise and encouragement during the examination session. Thus, the second factor appears to measure behaviors more specific to test sessions than the first factor.

Inattentiveness

The third and smallest factor, Inattentiveness, consists of 10 items and accounts for about 8% of the total variance. This factor is characterized by inadequate impulse control and attending behaviors. It, like the Avoidance factor, is associated with other overarching dimensions of child behavior. In the contexts of home and school, this factor is similar to one labeled Conduct Disorder by Quay (1986) and Externalizing by Achenbach and Edelbrock (1983).

CRITERION-RELATED VALIDITY

A substantial number of criterion-related validity studies are presented in the GATSB manual, including bivariate correlation analyses, canonical correlation analyses, and discriminant function analyses. However, other studies were needed. Since the publication of GATSB, researchers have used it to investigate whether children's test behaviors are affected by criterion-related bias (Glutting, Oakland, & Konold, 1994), item bias (Nandakumar, Glutting, & Oakland, 1993), and factor bias (Konold, Glutting, Oakland, & O'Donnell, 1995). The criterion-related validity of the GATSB also has been examined for children with attention deficit–hyperactivity disorders (Glutting, Robins, & deLancy, 1997), and for samples of both normal and referred children (Glutting et al., 1996).

Findings from the previously cited investigations strongly attest to the GATSB's validity for use with children who differ by age, gender, race, and ethnicity. The correlation between the GATSB Total Score and the WISC-III Full Scale IQ for children generally is substantial: −.36. This correlation is −.28 for Anglos, −.37 for African-Americans, and −.55 for Hispanics. Thus, for children generally, approximately 13% of the variance associated with Full Scale scores may be attributable to their test-taking behaviors. The amount of variance attributable to test-taking behaviors is 14% for African-American children and 30% for Hispanic children. Among the three factors, Avoidance accounts for the plurality of the variance.

The correlation between the GATSB Total Score and the WIAT Total Composite score again is substantial: −.38. Thus, approximately 14% of the variance associated with the measurement of achievement may be attributable to their test-taking behaviors. Correlations between the GATSB Total Score and the four WIAT Composite scores are similar.

Perhaps the most interesting finding from these studies is that children with inappropriate test behaviors, as measured by the GATSB, obtain WISC-III Full Scale IQs anywhere from 7 to 10 points lower than children with more suitable test behaviors. Effect sizes this large (more than half of a standard deviation) represent a substantial difference in IQs and testifies to the importance of observing peripheral test behaviors as a means of validating the integrity of formal scores obtained on standardized tests of intelligence and achievement.

RECENT CRITERION-RELATED FINDINGS

The intrasession and exosession validity of the GATSB was reexamined using new data. Data were obtained in 1995 and 2003 on 224 children and
adolescents (67% males) who ranged in age from 6 years, 0 months through 16 years, 11 months (M = 10.7 years, SD = 2.5 years). Participants were public school students from 34 school districts.

The sample was heterogeneous with respect to race and ethnicity. Although 27% attended regular education, the majority were eligible for special education: 48% were identified as having a learning disability, 11% with mental retardation, 8% with a serious emotional disturbance, 2% with an attention deficit disorder, and the remaining 5% with other impairments. Examiners characteristics were recorded as well (N = 34). Postgraduate experience among the 34 examiners ranged from 7 to 21 years (M = 12.7 years, SD = 7.1 years).

Standardized scales were used as validity criteria. Intrasession validity was assessed through WISC-III Full Scale IQs (FSIQ), Verbal Scale IQs (VIQ), and Performance Scale IQs (PIQ). Exosession validity (i.e., the generalizability of observations from the test session to other settings) was evaluated through teacher ratings on the Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993). The ASCA has six core scales: Attention Deficit–Hyperactivity, Solitary Aggressive (Provocative), Solitary Aggressive (Impulsive), Oppositional Defiant, Diffident, and Avoidant. The ASCA also yields two overall dimensions: Overreactivity (obtained by adding item scores from the first four core scales) and Underreactivity (based on item scores from the last two core scales). All eight ASCA scores were used as criteria.

Table 14.1 shows statistics for the GATSB predictors and criteria. Means for the GATSB predictors were near their population expectancy. Alternatively, means for the WISC-III criteria were below their population averages, while means for the Overreactivity, Attention Deficit–Hyperactivity, and Oppositional Defiant were elevated (i.e., suggesting more problem behaviors). These departures were anticipated given that the majority of the sample was eligible for special education services. The departures also are comparable to levels reported for a referral cohort during an earlier investigation of the GATSB (Glutting et al., 1996).

Correlations between the GATSB Total Score and FSIQs, VIQs, and PIQs are similar and significant (p < .001): −.29, −.26, and −.29, respectively (Table 14.2). Scores from the GATSB Avoidance scale are more predictive than those from the Inattentiveness and Uncooperative Mood scales. The correlations between the GATSB Total Score and IQs are somewhat lower than that found when norming the GATSB (i.e., the correlation between the GATSB Total Score and the WISC-III Full Scale IQ was −.36). The somewhat lower correlations may be due to this sample being smaller and unrepresentative of children nationally.

Whether test-taking behaviors impact intelligence test scores constitutes the more practical and important question. This question was examined by dividing children according to whether they exhibited compliant or non-

---

**Table 14.1** Distribution Statistics of GATSB Predictors and IQ, Achievement, and Behavioral Criteria

<table>
<thead>
<tr>
<th>GATSB Predictor</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Score</td>
<td>52.0</td>
<td>8.5</td>
</tr>
<tr>
<td>Avoidance</td>
<td>52.8</td>
<td>9.8</td>
</tr>
<tr>
<td>Inattentiveness</td>
<td>51.0</td>
<td>8.5</td>
</tr>
<tr>
<td>Uncooperative Mood</td>
<td>50.5</td>
<td>7.4</td>
</tr>
</tbody>
</table>

**WISC-III Criteria**

<table>
<thead>
<tr>
<th>FSIQ</th>
<th>−.29&lt;sup&gt;a&lt;/sup&gt;</th>
<th>−.41&lt;sup&gt;a&lt;/sup&gt;</th>
<th>−.10</th>
<th>−.13&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIQ</td>
<td>−.26&lt;sup&gt;a&lt;/sup&gt;</td>
<td>−.37&lt;sup&gt;a&lt;/sup&gt;</td>
<td>−.08</td>
<td>−.12&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>PIQ</td>
<td>−.29&lt;sup&gt;a&lt;/sup&gt;</td>
<td>−.41&lt;sup&gt;a&lt;/sup&gt;</td>
<td>−.09</td>
<td>−.12&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Note: FSIQ = Full Scale IQ, VIQ = Verbal Scale IQ, PIQ = Performance Scale IQ, N = 224.

<sup>a</sup>p < .001.
<sup>b</sup>p < .05.
compliant test behaviors during administrations of the WISC-III (Table 14.3). The former group had GATSB Total Scores in the average range (i.e., T scores < 59), while the latter group had GATSB Total Scores 1 standard deviation or more above the mean (i.e., T scores > 60). In contrast to children with compliant test behaviors, those who did not demonstrate compliance averaged 8.6 points lower FSIQs, 7.2 points lower VIQs, and 9.5 points lower PIQs. Thus, children with compliant test behavior obtain FSIQs, VIQs, and PIQs 0.5 to .66 of a standard deviation higher than children with less suitable test behaviors.

These data confirm that children’s test-taking behaviors are meaningfully related to the magnitude of their measured intelligence. Current findings generally are consistent with earlier findings and support the conclusion that test observations are modestly yet meaningfully related to children’s IQs.

The GATSB and ASCA evaluate similar constructs albeit displayed in different contexts. The GATSB assesses behaviors displayed in the context of testing while the ASCA assesses behaviors displayed in the context of school. Construct validity is suggested when an appropriate pattern of convergent and divergent associations is found between similar tests (Campbell, 1960; Thorndike, 1982). Higher correlations were expected between identical or convergent scales from the two tests (e.g., GATSB Avoidance and ASCA Avoidance) and lower correlations were expected between divergent scales (e.g., GATSB Avoidance and ASCA Attention Deficit-Hyperactivity). As expected, the anticipated convergent associations were collectively higher than divergent associations and support inferences of the GATSB’s construct validity (Table 14.4). Thus, behaviors observed in test sessions are more related to behaviors in the same than in other contexts.

The averaged coefficient between the GATSB and ASCA provides evidence of the degree test behaviors are generalizable across contexts to situations important to child adjustment, namely their behavior in school.

### Table 14.4 Correlations of GATSB Predictors and ASCA

<table>
<thead>
<tr>
<th>ASCA Criteria</th>
<th>Total Score</th>
<th>Avoidance</th>
<th>Inattentiveness</th>
<th>Uncooperative Mood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underreactivity</td>
<td>.24**</td>
<td>.38*</td>
<td>.07</td>
<td>.04</td>
</tr>
<tr>
<td>Overreactivity</td>
<td>.18*</td>
<td>.07</td>
<td>.22*</td>
<td>.18**</td>
</tr>
<tr>
<td>Avoidant</td>
<td>.17*</td>
<td>.23*</td>
<td>.10</td>
<td>.03</td>
</tr>
<tr>
<td>Diffident</td>
<td>.12</td>
<td>.28*</td>
<td>-.04</td>
<td>-.01</td>
</tr>
<tr>
<td>Solitary Aggressive-Provocative</td>
<td>.15*</td>
<td>.09</td>
<td>.19*</td>
<td>.15*</td>
</tr>
<tr>
<td>Solitary Aggressive-Impulsive</td>
<td>.21*</td>
<td>.09</td>
<td>.29*</td>
<td>.21*</td>
</tr>
<tr>
<td>Attention Deficit-Hyperactivity</td>
<td>.15*</td>
<td>.05</td>
<td>.23*</td>
<td>.21*</td>
</tr>
<tr>
<td>Oppositional Defiant</td>
<td>.07</td>
<td>.08</td>
<td>.09</td>
<td>.01</td>
</tr>
</tbody>
</table>

Note: N = 224.

*p < .001***

*p < .01.

*p < .05.

The averaged coefficient among correlations between the GATSB and ASCA is low (r = .14) and lower than the average correlations between GATSB Total Score and FS, V, and PIQs (r = .28). Furthermore, the averaged exosession validity coefficient of .14 suggests that up to 98% of the variation in test behaviors is accounted for by in-school behaviors measured by the ASCA.

The belief that behaviors psychologists observe while testing may not generalize well to behaviors reported by teachers and parents also is supported by finding for 122 children ages 6 through 16 referred for a psychological evaluation (Daleiden, Brabman, & Benton, 2002). The GATSB correlated moderately with the WISC-III and Woodcock-Johnson Psychoeducational Battery—Revised and the Wide Range Assessment of Memory and Learning and correlated considerably lower with reports from parents and teachers on the Child Behavior Checklist. The conclusion that most of the variation in test behaviors may be specific to the context in which they occur seems reasonable. However, clinicians are advised to confer with teachers, parents, and other significant sources to determine the generalizability of findings from measures of test-taking abilities to broader contexts on a case-by-case basis. Moreover, knowledge of behaviors displayed at school and home is essential when interpreting GATSB data.

### POSSIBLE USES OF TEST BEHAVIOR INFORMATION

Accomplished examiners typically rely heavily on their observations to assist them in various components of the assessment process. As previously

---

**Table 14.3 Differences in WISC-III Scores for Children Showing Compliant and Noncompliant Test Behaviors**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>Mean</th>
<th>t-value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSIQ</td>
<td>Compliant (n = 182)</td>
<td>87.7</td>
<td>2.81</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>Noncompliant (n = 42)</td>
<td>79.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIQ</td>
<td>Compliant (n = 182)</td>
<td>86.2</td>
<td>2.35</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>Noncompliant (n = 42)</td>
<td>79.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIQ</td>
<td>Compliant (n = 182)</td>
<td>91.7</td>
<td>3.09</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td>Noncompliant (n = 42)</td>
<td>82.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: FSIQ = Full Scale IQ, VIQ = Verbal Scale IQ, PIQ = Performance Scale IQ, N = 224.
noted, this information assists them in understanding the processes used by children in arriving at their responses, in screening important qualities not measured directly by the test, in discussing the child’s qualities intelligibly in conferences with parents and teachers, and in making needed modifications in how the test is administered. Knowledge of test behaviors also enables examiners to decide whether the test results accurately reflect the child’s cognitive abilities.

A RETURN TO TWO-TEST PERFORMANCE STANDARDS

One of two standards typically is used when evaluating whether the test data are considered to be valid: optimal and typical performance. Examiners often differ in their views as to which should be used.

OPTIMAL PERFORMANCE

Examiners favoring optimal performance strive to create conditions that facilitate the examinee’s highest performance (for tests of ability) or most representative performance (for tests of personality and social skills) as the standard for test performance. Test behaviors that negatively influence test performance and thus attenuate or in other ways distort scores are thought to be inappropriate. Those who employ optimal performance standards generally believe test scores from examinees who display unmotivated, uncooperative, and inattentive test behaviors are invalid and thus should be discarded.

TYPICAL PERFORMANCE

Examiners favoring typical performance standards also are likely to interpret these test behaviors as being inappropriate and negatively influencing test performance. However, when assessing children whose unsuitable test-session behaviors are similar to those displayed at home and school, those using typical performance standards are likely to interpret aberrant test-taking behaviors as reflecting broader personal traits and thus will accept the test results as being valid, given their belief that the deleterious behaviors observed during the test also are likely to negatively influence the examinee’s general behaviors. Thus, the results from cognitive tests are thought to reflect real-life conditions and are valid. In contrast, those who adhere to an optimal performance standard are more inclined to reject the data as being invalid.

The following case study, taken from the GATSB manual, demonstrates how the GATSB data may be used to facilitate the assessment of cognitive abilities.

Anne, a 13-year-old middle-school student, was seen for testing and a clinical interview on one occasion, from 8:30 to 11:45 AM. She was suitably dressed and groomed, wearing tennis shoes, modest jewelry, and a matching skirt and blouse. Her short hair was clean and informally styled. Anne is of average height and somewhat overweight. School records indicate normal visual and auditory acuity, and her primary and only language is English. School attendance is regular. Anne indicated an absence of medical problems; moreover, no problems were apparent in her fine- and gross-motor coordination skills and in her linguistic proficiency. Anne had been informed about and was fully oriented to the testing; however, she expressed concern about being tested. As indicated in the following GATSB results, her test-taking behaviors, assessed after administration of the WISC were problematic.

<table>
<thead>
<tr>
<th>Score</th>
<th>Confidence Interval</th>
<th>Cumulative Percentile Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Score</td>
<td>74</td>
<td>± 2.83</td>
</tr>
<tr>
<td>Avoidance Scale</td>
<td>76</td>
<td>± 3.74</td>
</tr>
<tr>
<td>Inattentiveness Scale</td>
<td>84</td>
<td>± 4.80</td>
</tr>
<tr>
<td>Uncooperative Mood Scale</td>
<td>59</td>
<td>± 3.46</td>
</tr>
</tbody>
</table>

Anne’s Total Score T score of 74 reflects an atypical pattern of test session behavior. This atypical pattern is also reflected in the highly elevated scores on the Avoidance (T score = 76) and Inattentiveness (T score = 84) scales. The highly elevated Avoidance score reflects Anne’s inability to remember test directions, her lack of eye contact with the examiner, her frequent requests to take a break, and a deterioration in performance during the last three WISC subtests that was a result of her withdrawal.

Anne’s highly elevated score on the Inattentiveness scale reflects her difficulty in completing work within time limits and her failure to attend to test directions. It also reflects her increasing restlessness during the test session and the deterioration in her efforts toward the end of the session.

Discussions with Anne’s English teacher and counselor after the testing confirmed the presence of the inappropriate behaviors in class. Thus, there is
some indication that the behaviors observed during the WISC-III administration are situational. The examiner conducting this evaluation uses optimal performance as a standard. Consequently, Anne’s WISC results are not reported. The examiner will confer with Anne’s parents and Anne to discuss the extent to which these and similar behaviors may be displayed at home. After these conferences the examiner will work with Anne to minimize any adverse effects avoidance and inattentive behaviors have on a second administration of the WISC.

CONCLUSION

Progress is being made in our ability to carefully observe and understand the relevance of test-taking behaviors of children and youth. Research during the last two decades has helped define the nature of important test-taking behaviors, to measure them reliably, and to understand their degree of impact on measures of cognitive abilities. Most recently developed individually administered tests of cognitive abilities include methods to assess test-taking behaviors—albeit using methods that lack norms and demonstrated empirical validity. Continued progress in improving the use of these observational abilities requires added emphasis in two areas: instruction and research.

In reference to instruction, most clinicians have little formal study on issues important to observing, recording, and interpreting test-taking behaviors. As a result, clinicians generally rely on time-honored but untested informal methods when recording and interpreting test-taking behaviors. Our profession now has available a new and growing body of information that leads to the conclusion that the use of standardized and norm-referenced measures of test-taking behaviors can significantly enhance assessment. Professors teaching assessment courses should include this content when preparing student clinicians. Established professionals should become acquainted with this content in other ways.

Research on test-taking behaviors began about 20 years ago and thus is in its infancy. Additional research is needed to help verify the full range of behaviors that may characterize test-taking behaviors, to fully define this construct, and to continue investigations as to possible demographic differences in these qualities. Relationships between measures of test-taking behaviors (e.g., the GATSB) and performance on measures other than the WISC and WIAT are greatly needed. We concur with the Kaufmans (1994) that information from the GATSB may be relevant to the understanding of a wide range of measures. These may include the Clinical Evaluation of Language Functions—Third Edition (Semel, Wiig, & Secord 1995), Differential Abilities Scale (Elliott, 1995), Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1985), Stanford-Binet Intelligence Scale: Fifth Edition (Roid, 2003), Woodcock-Johnson Tests of Cognitive Abilities (Woodcock, McGrew, & Mather, 2001), and various other individually administered measures. In addition, the assets and limitations of test behaviors to our understanding of noncognitive qualities (e.g., personality and temperament) deserve further study.

REFERENCES


