DOCUMENT RESUME

ED 453 950 PS 029 533

AUTHOR Grigorenko, Elena L.; Sternberg, Robert J.
TITLE Assessing Cognitive Development in Early Childhood. Early Childhood Development.
INSTITUTION World Bank, Washington, DC. Human Development Network.
PUB DATE 1999-12-00
NOTE 68p.
AVAILABLE FROM World Bank, 1818 H Street, N.W., Washington, DC 20433; Tel: 202-473-3427; Fax: 202-522-3233; e-mail: myoung3@worldbank.org.
PUB TYPE Information Analyses (070)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Behavior Rating Scales; *Cognitive Tests; Criterion Referenced Tests; Infants; Measures (Individuals); Norm Referenced Tests; Preschool Children; *Preschool Tests; *Psychological Testing; *Psychometrics; Screening Tests; *Standardized Tests; Test Reliability; *Test Selection; Test Validity; Theories; Toddlers

ABSTRACT

Noting that the last 40 years have witnessed an enormous increase in the number of psychological tests designed for the assessment of competencies in very young children, this review summarizes the quantitative and qualitative characteristics of psychological tests and other assessment instruments used to evaluate the cognitive functioning of infants, toddlers, and preschool children. The review is presented in three parts. Part 1 summarizes general principles of early childhood assessment. Part 2 describes the major domains in which the various assessment tools can be compared, evaluated, and selected. Part 3 presents brief descriptions and evaluations of selected instruments. Appended is a list of 313 additional references regarding specific tests or the assessment process. (Contains 103 references.) (KB)

Early Child Development

ASSESSING COGNITIVE DEVELOPMENT IN EARLY CHILDHOOD

Elena L. Grigorenko

and

Robert J. Sternberg

December 1999

EDUCATION
THE WORLD BANK

ASSESSING COGNITIVE DEVELOPMENT IN EARLY CHILDHOOD

Elena L. Grigorenko

Yale University and Moscow State University

Robert J. Sternberg

Yale University

December 1999

Table of Contents

Introduction 1

I. General Principles of Early Childhood Assessment 1
- Testing 3
- Interviewing 3
- Behavioral Observations 3
- Synthesis 4

II. Factors Important for Evaluating and Selecting Tests 5
- Purpose 5
- Data 5
- Standardization 8
- Psychometric Properties 8
- Test Construction, Format, and Administration 10

III. Specific Early Childhood Tests 11
- Multidomain Assessment 11
- Standardized Tests 11
  - The Brazelton National Behavioral Assessment Scale (2nd Ed.) 11
  - The Bayley Scales of Infant Development—II 12
  - The NEPSY 13
  - The Griffiths Mental Development Scales 14
  - The Battelle Development Inventory 14
  - The Mullen Scales of Early Learning 15
  - The Peabody Picture Vocabulary Test-Revised 15
- Behavior Rating Scales 16
  - Childhood Behavior Checklist 13
  - The Conners' Behavior Rating Scales 13
  - The Vineland Adaptive Behavior Scales 13
- Theories-of-Cognition-Based Tests 18
  - The Wechsler Preschool and Primary Scale of Intelligence—Revised 13
  - The Wechsler Intelligence Scale for Children—Third Edition 13
  - The Stanford-Binet Intelligence Scale: Fourth Edition 13
  - The Cattell Infant Intelligence Scale 13
  - The McCarthy Scales of Children's Abilities 13
  - The Differential Ability Scales 13
  - The Woodcock-Johnson Psycho-Educational Battery-Revised: Tests of Cognitive Ability 13
  - The Kaufman Assessment Battery for Children 13
  - The Standard Raven Progressive Matrices 13

Criterion/Norm-Referenced Assessment 27
Norm-referenced comparison 13
The Brigance Diagnostics Inventory of Early Development—Revised 13
The Infant Psychological Developmental Scale 13
The Metropolitan Readiness Test 13
The Peabody Individual Achievement Test-Revised 13
Screening Devices 28
The Denver Developmental Screening Test—II 13
Other tests 13
The Miller Assessment for Preschoolers 13

IV. Concluding Remarks 30

References 31

Table 1 Major Domains of Young Children's Behavior and Tests Utilizable for Assessment of These Domains 2
Table 2 6

Introduction

Programs designed to improve the health, nutritional, or cognitive status of preschool children promise to take young children at risk and potentially to change their lives. Children whose opportunities in life otherwise might be severely diminished are given a chance to have a brighter start and a brighter future. But intervention programs inevitably vary in quality and impact. Investment of huge amounts of time, effort, and resources do not guarantee that such programs will produce the outcomes that are intended, particularly with regard to cognitive development. For this reason, program developers or those who fund them recognize the need to evaluate the health, nutritional, and psychological impact of such programs. In this review, we deal only with cognitive impact.

The last 40 years of psychological assessment have witnessed an enormous increase in the number of psychological tests designed for the assessment of competencies in infants, toddlers, and preschool children. It is fair to say that, today, early childhood assessment constitutes a growing field with new instruments being developed regularly. The purpose of this review is to summarize the quantitative and qualitative characteristics of psychological tests and other assessment instruments used to evaluate the cognitive functioning of infants, toddlers, and preschool children.

The review is divided into three parts. The first part summarizes general principles of early childhood assessment. The second part describes the major domains in which the various assessment tools can be compared, evaluated, and selected. Finally, the third part presents brief descriptions and evaluations of selected instruments.

I. General Principles of Early Childhood Assessment

Early childhood assessment is guided by five major principles.

First, no single test can address all questions or solve all problems. Thus, assessment of preschoolers ideally should rely on specific instruments for specific situations. Moreover, multiple instruments ought to be used, if possible, because no one instrument is likely to provide a complete assessment of all intended outcomes.

Second, there is a greater likelihood of substantial levels of measurement error in early childhood assessment than in assessment at any other period of a child's life. Therefore, accurate childhood assessment ideally requires that information from a highly structured assessment be integrated with information from other *types* of more semi-structured, open-ended assessments, such as interviews and behavioral observations. The combination of more and less structured assessments raises the probability that a complete picture of the data will emerge.

Third, young children usually perform better in the company of familiar adults. Thus, assessment of young children should be viewed as a collaborative enterprise between the assessor and the child's parents or caregivers. The more standard model of assessment in a sterile environment (such as a bare room with no other people besides the tester and test-taker) does not apply as well to young children as it does to older ones.

Fourth, young children's functioning is profoundly influenced by the setting in which the assessment takes place. The implication for assessment is that it is best to observe the child in a range of natural

settings, most importantly, the home and the childcare situation. Just as some adults function very differently in the home versus the workplace, all the more do many young children function differently in one setting versus another.

Fifth and finally, assessment should extend over a period of time (for developmental and general/mental-health tasks, 4 to 8 weeks is ideal) in order for assessors to gain a detailed understanding of the prevailing emotional themes, the range of functioning in the child and the caregivers, the degree of variation in the quality of the child's primary relationships, and the relative influence of situational factors, such as family circumstances and chronic stressors.

To realize these principles of early childhood assessment, skillful evaluators (1) form alliances with caregivers, (2) utilize structured and semi-structured interviewing techniques, (3) ask questions to clarify but not disrupt the caregivers' accounts, (4) listen to participants and observe the affect they demonstrate as well as the content they provide, and (5) guide both children and caregivers through the assessment process.

Knoff (1999) emphasized that assessment should be multi-method, multi-source, and multi-setting. Among the many techniques used in early childhood assessment, the three most prevalent ones are testing, interviewing, and behavioral observations.

Testing

Testing is an assessment technique that employs standardized instruments. For an assessment instrument to be called a test, it should (a) have a clear purpose (e.g., target a certain psychological domain or multiple domains), (b) provide explicit methods to evaluate data (e.g., specify correct and incorrect responses), (c) rely on a standardization scheme (e.g., link individual data to population data), (d) meet certain psychometric criteria, and (e) have appropriate test format, construction, and administration.

Interviewing

Because much of the information about young children's daily functioning is best delivered by caregivers, interviews with primary caregivers are often central to a comprehensive developmental assessment. The main purpose of interviews with caregivers is to gather information about the child's developmental history and the caregivers' perceptions of the child's level of functioning. The important areas to cover are (1) the history of the mother's pregnancy, delivery, and immediate perinatal period; (2) the child's medical history; (3) the child's developmental milestones; (4) the number, ages, and health of family members; (5) the infant's fit in the family's daily life; (6) each parent's interpretation of the significance of the child to their lives and the life of the family; and (6) the child's functioning in several areas.

The following aspects of the child's development and well-being should be assessed directly and evaluated in parental interviews: (a) motor development; (b) general activity level; (c) speech and communication; (d) problem solving; (e) play; (f) self-regulation (e.g., independence, initiative, need for routines); and (g) relationships with others, including level of social responsiveness.

Behavioral Observations

Assessment of young children should include general descriptions of the children's behavior and qualitative accounts of the children's behavior in at least one structured setting. Special areas of interest are the children's (a) responses to developmental tasks (excitement, positive versus negative affect, energy versus lack of energy, quickness versus slowness, and deliberation versus impulsiveness); (b) ability to cope with frustration; (c) engagement with the adult world; (d) range of emotional expressiveness; and (e) capacity for persistence and sustained attention. Behavioral observations require a mixture of free-floating attention to the child's behavior and the more focused attention to specific situational responses that is inherent in any structured or semi-structured assessment. In other words, the assessor should be attentive and sensitive to whatever occurs, but she or he should also have a mental plan, a mental map that serves as a framework for organizing observations collected during the assessment. Key components to be addressed in such a framework are (a) the quality of the evaluative environment or "evaluative atmosphere" and the affective attitudes of the caregiver and the child; (b) situational involvement (curiosity and interest versus detachment and lack of interest), (c) engagement of others (the child's interactions with the caregiver and the examiner, the caregiver's involvement with the child), and (d) reaction to change (initial greeting, ending of assessment, transitional periods from task to task, and so on).

There are at minimum three levels at which the assessor needs to collect her or his information. The first and most apparent is the level at which the child's reactions and responses to the structured assessment items are registered. The data collected at this level should not be solely confined to whether or not the child passes or fails a given task but also should address qualitatively how the child approaches and deals with the task. The second level of observation concerns the child's reaction to the assessment situation, apart from the formal tasks. The assessor must register whether the child approaches toys, initiates interactions, refers to the examiner or to his or her caregiver, reacts to the assessor and the situation at the beginning and the end of the evaluation session, and so on. The third level of observation addresses the interaction between caregiver and infant. Observations at this level are made throughout the assessment process, and various fluctuations in these interactions are registered.

Here are some points of observation for judging caregiver-child interactions:

1. Does the child appeal to the caregiver for help and reassurance?
2. Does the child call his or her success to the attention of the caregiver?
3. Does the caregiver respond to the child's success/failure?
4. How does the caregiver hold and soothe her or his child?
5. Does the child distance him or herself from the caregiver to work with tasks or to explore?
6. Does the caregiver show his or her involvement? Is the involvement intrusive or encouraging?
7. Does the caregiver comfortably assist the child?
8. Is the caregiver withdrawn?

Synthesis

The synthesis involved in the assessment of young children involves summarizing both qualitative and quantitative data gathered from interviews, observations, and testing. Whether assessment has been carried out for clinical or research purposes, it is important to keep in mind that there is no single instrument or technique that can capture the full variability of the child's performance in a variety of settings. In working with young children, the utilization of the multi-trait—multi-method approach is therefore a must.

II. Factors Important for Evaluating and Selecting Tests

As has been stressed above, the main rule of early childhood assessment is to use a variety of tests. Assessors working with young children should not fetter themselves by knowing about only a few tests. As the old adage goes, "If all you have is a hammer, all your problems start looking like nails." Table 1 lists major domains of young children's behavior and suggests tests that can be used for each specific domain.

Five distinct issues should be considered when evaluating a developmental test: (1) the extent to which the purpose of the test fits the purpose of the assessment, (2) the source of the test data, (3) the quality of the test standardization and the relevance of the population on which it was standardized to the population being tested, (4) the psychometric properties of the test, and (5) the qualitative characteristics of the test (i.e., test construction, format, and administration). Criteria for evaluating tests for young children are summarized in Table 2.

Purpose

Early childhood tests are traditionally used for various purposes, among which are diagnosis, screening, intervention planning, and research. Typically, there are different tests available for these different purposes, but sometimes the same test can be used for different purposes.

In general, tests administered for diagnostic purposes are administered individually, with the goal of obtaining a comprehensive picture of the child's functioning in a number of areas. Screening tests are used primarily when many children are to be assessed and when evaluators desire a relatively brief, cheap, and user-friendly instrument that will allow them to identify children, especially infants, who may be "at-risk" for developmental delay. Such tests also may be used to evaluate the impact of an intervention program. Testing (both individual and group) also can be aimed at identifying objectives and steps for individualized intervention programs. Moreover, individual/group testing can be used to track children's achievement of desired goals over time and to monitor the effects of intervention programs. Finally, testing carried out for research purposes may be also designed for individual or group administration.

Data

The main purpose of administering formal tests to children is to collect and organize data regarding children's level of functioning. In working with young children, the most salient types of data are those obtained from direct assessment, observation, and caregiver interviews (reports). Most developmental tests utilize some combination of all three of these types of data; there are, however, specialized instruments capitalizing on a single type of data. Each of the three types of data has its strengths and weaknesses. For example, direct assessment typically utilizes standardized types of administration, which readily allow a particular child's performance to be compared to the performance of other children. It is crucial that the standardization data be relevant to the children being assessed. Yet, standardized test items typically open only a small window into a child's life, and a child's performance can be greatly influenced by the child's motivation, mood, comfort, and rapport with the assessor. It is therefore important not to draw conclusions from limited test data that go beyond what the data truly can say. Observational data, too, have their advantages and limitations. When collected in naturalistic settings,

Table 1. Major Domains of Young Children's Behavior and Tests Utilizable for Assessments of These Domains

Targeted Domains	Testing
General Cognitive Abilities	Wechsler Preschool and Primary Test of Intelligence-Revised (WPPSI-R) Differential Ability Scale-Preschool Core (DAS) Stanford-Binet Intelligence Scale: Fourth Edition (SBIS: IV) McCarthy Scales of Children's Abilities Kaufman Assessment Battery for Children (K-ABC) Woodcock-Johnson Psychoeducational Test Battery-Revised (WJ-R) Mullen Scales of Early Learning (MSEL)
Language and Language Related Processes	Peabody Picture Vocabulary Test-Third Edition (PPVT-III) Verbal subtest of standardized intellectual batteries Preschool Language Scale-Third Edition (PSL-3) Test of Language Development-Second Edition (TOLD-2) Vineland Adaptive Behavior Scale-Communication Domain NEPSY (Language Domain) MSEL
Nonverbal Processing	K-ABC Simultaneous Scale Nonverbal subtests of DAS, WPPSI-R, McCarthy, BSID:IV
Motor	Purdue Pegboard Vineland Motor Domain McCarthy Motor Scale NEPSY (Fingertip Tapping, Imitating Hand Positions, Manual Motor Sequences, Finger Discrimination) MSEL
Executive Functions	WPPSI-R Animal Pegs NEPSY (Tower, Statue, Knock and Tap)
Memory	DAS (Recall of Digits, Recall of Objects) McCarthy Memory Scale BISD:IV (Bead Memory, Sentence Memory) K-ABC (Face Recognition, Hand Movements, Number Recall, Word Order, Spatial Memory) NEPSY (Memory for Faces, Memory for Names, Narrative Memory, Sentence Repetition, List Learning)
Social/Emotional Adjustment	Vineland Socialization Domain The Child Behavior Checklist Conners' Behavior Rating Scales
Preacademic Skills	K-ABC (Achievement Scale) WJ-R WPPSI-R (Arithmetic, Information) MSEL DAS (Matching Letter Like Forms, Early Number Concept)

Table 2. Criteria for Evaluating Technical Characteristics of Early Childhood Cognitive Competence Assessment Devices[1]

Criterion	Specifications	Evaluations
Purpose	1. Diagnostic 2. Screening 3. Intervention planning 4. Research	
Data	1. Direct assessment 2. Observation 3. Caregiver (parent/teacher) interview/report	
Standardizability	1. Standardization	
	(1) Size of normative group	
	N = 200 per each 1-year interval and N ≥ 2,000 overall	Good
	N = 100 per each 1-year interval and N ≥ 1,000 overall	Adequate
	Neither requirement above is met	Inadequate
	(2) Satisfactoriness of normative data	
	Collected in 1988 or later	Good
	Collected between 1978 and 1987	Adequate
	Collected in 1977 or earlier	Inadequate
	(3) Representativeness of the general population by the normative sample	
	Normative sample represents the general population on ≥ 5 important demographic variables (e.g., gender, nationality) with SES included	Good
	Normative sample represents the general population on ≥ 3 important demographic variables with SES included	Adequate
	Neither criterion is met	Inadequate
	2. Norm table age stratification	
	1-2 months	Good
	3-4 months	Adequate
	> 4 months	Inadequate
Psychometric Properties	1. Reliability	
	(1) Internal consistency reliability coefficient	
	≥ .90	Good
	.80-.89	Adequate
	<.80	Inadequate
	(2) Test-retest reliability coefficient	
	≥ .90	Good
	.80-.89	Adequate
	<.80	Inadequate
		(continued)

[1] Adapted from Algonso & Flanagan, 1999.

Table 2. Criteria for Evaluating Technical Characteristics of Early Childhood Cognitive Competence Assessment Devices *(Continued)*

Criterion	Specifications	Evaluations
Psychometric Properties	(3) Sample size and representativeness of test-retest sample	
	N ≥ 100 and represents the general population on ≥ 5 important demographic variables	Good
	N ≥ 50 and represents the general population on ≥ 3 important demographic variables	Adequate
	Neither criterion is met	Inadequate
Inadequate		
	(4) Age range of the test-retest sample	
	≤ 1-year interval	Good
	≤ 2-year interval	Adequate
	≥ 2-year interval (or extends beyond the preschool age range (i.e., 3 to 5 years)	Inadequate
	(5) Length of test-retest interval	
	≤ 3 months	Good
	≥ 3, but ≤ 6 months	Adequate
	≥ 6 months	Inadequate
	2. Validity (content, criterion-related, and construct)	
	3 types of validity evaluated	Good
	2 types of validity evaluated	Adequate
	1 type of validity evaluated	Inadequate
	3. Floors	
	Raw score of 1 is associated with a standard score > 2 sd(s) below the normative mean	Adequate
	Raw score of 1 is associated with a standard score ≤ 2 sd(s) below the normative mean	Inadequate
	4. Item Gradients	
	No item gradient violations occur or all item gradient violations are between 2 and 3 sd(s) below the normative mean	Good
	All item gradient violations occur between 1 and 3 sd(s) below the normative mean	Adequate
	All or any portio of item gradient violations occur between the mean and 1 sd below the normative mean	Inadequate

mula). Coefficient alpha refers to the corrected correlation between all possible pairings of split-halves of the test.

Validity refers to the extent to which a test measures what it is supposed to measure. There are several different types of validity that are useful in determining the overall quality of a test. Tests are expected to demonstrate face, content, criterion-related, and construct validity. Face validity refers to the extent to which the test appears to examinees to measure what it is supposed to measure. Face validity is

observations can counter the "artificiality" of direct assessment, providing, presumably, more ecologically valid data. However, in naturalistic observation, the relative contributions of the child versus the context cannot readily be partitioned. Specifically, based on observational data, it is sometimes unclear whether the child's observed behavior is due to a problem within the child or to a problem in the environment in which the child was observed. Similarly, caregivers' reports may be useful for screening purposes, for detecting rare problem behaviors, for obtaining information regarding behavior that may be difficult to elicit in the structured assessment, and for assessing the caregivers' perspectives of the children. But such reports bear a stamp of subjectivity and, therefore, should be viewed cautiously (Meisels & Waskik, 1990).

Standardization

The standardization of data refers to the availability of normative data regarding children's typical performance on a given test. Normative data allow an experienced assessor to determine a specific child's placement relative to the normative group. This relative placement is usually expressed in terms of percentile ranks or standard scores.

The critical issue here is to determine whether a given test's standardization sample is representative of the population of children the test user plans to assess. The norms should be appropriate for the children's specific historical context, locality, gender, ethnicity, and economic status. For example, U.S. norms applied to African data will be of little use. Moreover, it is important to take into account that the same item presents different challenges to children in different populations, so that the standardization data may not be meaningful when transferred from one population to another. For example, difficult vocabulary may be less relevant in the lives of some children than in the lives of others. Some children may never encounter these words in their lives, whereas other children may encounter them with some frequency.

Psychometric Properties

The psychometric properties of psychological instruments assessing characteristics of early development are crucial because (1) young children's abilities change rapidly and instruments have to be sensitive to these subtle changes, (2) many traditional instruments represent downward extensions of tests that originally were designed for older children and, therefore, were not designed with the preschool child in mind; and (3) the use of traditional intelligence tests with young children has been criticized extensively due to poor psychometric properties of the tests (Alfonso & Flanagan, 1999).

The two key psychometric properties of a test are the test's reliability and validity. In general, reliability indicates the test's dependability (i.e., the test's ability to produce similar results under differing conditions). Specifically, (1) test-retest reliability indicates the test's temporal stability (as shown by the correlation between test scores obtained at one time with test scores of the same individuals obtained at a later time); (2) inter-rater reliability indicates the degree to which test scores are insensitive to individual differences between different assessors; and (3) internal consistency indicates the degree to which different items of the test measure the same underlying construct (Salvia & Ysseldyke, 1991). Internal consistency can be measured in a number of ways. Odd-even reliability refers to the correlation between scores on the odd and even numbered items of a test (corrected by the Spearman-Brown for-

important because examinees sometimes lack motivation to succeed on tests that do not appear to be face valid to the examinees. Content validity refers to the extent to which the test appears to experts to measure what it is supposed to measure. Content validity is important because, in its absence, experts may not accept the results of the test, regardless of its statistical properties. Criterion-related validity refers to the extent to which a test is correlated with other measures with which it is supposed to be correlated. Criterion-related validity can be either predictive or concurrent. Predictive validity refers to the extent to which the test predicts criterion performance to be obtained in the relatively distant future. Concurrent validity refers to the extent to which the test predicts criterion performance collected concurrently or in the very near future (such as the same day or the next day). Psychometricians also sometimes distinguish between convergent validity and discriminant validity. Convergent validity refers to the extent to which a test or test battery correlates with other measures with which it is supposed to correlate. Discriminant validity refers to the extent to which a test fails to correlate with other measures with which it is *not* supposed to correlate. For example, one would wish a test of intellectual ability to predict school performance but not to predict physical prowess.

A number of general standards for assessing the adequacy of developmental assessment instruments for preschoolers have been proposed by Bracken (1987; see also Flanagan & Alfonso, 1995).

First, assessment tools should possess adequate reliability, as evident by a median subtest internal-consistency reliability of at least .80 and total test internal-consistency and stability reliability coefficients of at least .90.

Second, developmental tests should have adequate "floor space." Specifically, tests should have enough low-level items so as to allow children to score at least two standard deviations below the mean on the overall score as well as on all subtest scores. Adequate tests floors are important for distinguishing children performing at different levels of functioning (average, low average, borderline, retarded). When tests floors are inadequate, they (a) provide scores that tend to overestimate the cognitive functioning of individuals with various degrees of retardation (i.e., mild, moderate, severe) and (b) provide more information about what a child cannot do than about what a child can do. (We would add that certain tests also should have adequate "ceiling space," with enough difficult items so as to allow children to score at least two standard deviations above the mean on the overall score as well as all subtest scores. In particular, tests of minimum competencies do not need ceiling space, but tests of the full range of competencies should allow such space.)

Third, the subtest item gradient should not be too steep. Specifically, each standard deviation of scores in the children's performance should consist of at least three raw score items. Put another way, large differences in standardized scores should not stem from small differences in actual raw-test scores. A subtest item gradient, by referring to the amount of change in a child's standard score that is associated with a one-unit change in his/her raw score, indicates the sensitivity of the test (i.e., the test's potential to detect fine gradations in cognitive performance within and across competency levels). If the subtest gradient is too steep, then strong conclusions may be made on the basis of raw score differences that reflect little more than chance fluctuations in the data. Inadequate test floors and item-gradient violations seriously jeopardize the quality of testing instruments, especially when the tests are used with young children.

Test Construction, Format, and Administration

The qualitative characteristics of tests for young children (test construction, format, and administration) include such characteristics as the appropriateness and attractiveness of testing material for young children, the opportunities for teaching to the tasks being tested, the comprehensiveness of the instructions, the appropriateness of the test for multicultural populations, and the testing environment.

In interpreting test results of children from diverse cultural backgrounds, Newland (1971) suggested placing various tests on the product-dominant/process-dominant continuum, where product-dominant tests depend on accumulated knowledge and environmental experiences to a greater extent than do process-dominant tests, which are assumed to assess fundamental learning and thinking processes.

III. Specific Early Childhood Tests

In order to balance breadth with depth, the following section provides a more extensive summary of some tests than of others. When an extensive summary is provided, the presentation of the test (a) provides a description of the test, (b) discusses the theory underlying the test, (c) summarizes qualitative and quantitative characteristics of the test, (d) addresses the purposes of the test and its reputation in the field, (e) comments on the test's strengths and weaknesses, (f) remarks on North American cultural biases that may be embedded in the test, and (g) comments on applications of the test in diverse countries.

Most of the existing infant/toddler/preschooler tests can be clustered into one of three categories, each based on a different theoretical and/or psychometric model (Gilliam & Mayes, in press). These groups are (1) multidomain development tests, (2) theories-of-cognition-based tests, and (3) criterion (norm)-referenced assessments.

Multidomain Assessment

The multidomain model is arguably both the most widely used and the oldest model of infant-toddler-preschooler assessment. The theoretical basis of the model is that child development is an interactively unfolding, continuous process that occurs in several distinct but interrelated domains (Gesell, 1940). Traditionally, these domains include (1) motor (fine and gross motor skills), (2) communication (receptive and expressive language), (3) cognition (problem-solving skills), (4) adaptive competence (self-help behaviors—dressing, eating, toileting, and so on), and (5) personal-social competence (social competence, emotional regulation, and sense of self). These domains are covered comprehensively in some modern assessment devices [e.g., the Bayley Scales of Infant Development (Bayley, 1993); the Battelle Developmental Inventory (Newborg, Stock, Wnek, et al., 1984); the Griffiths Mental Development Scales (Griffiths, 1954, 1979); the Mullen Scales of Early Learning (Mullen, 1995); and the NEPSY (Korkman, Kirk, Kemp, 1998)] and partially in others [e.g., the Receptive-Expressive Emergent Language Scale-2 (Bzoch & League, 1991); the Peabody Developmental Motor Scale (Folio & Fewell, 1983); and the Vineland Social-Emotional Early Childhood Scales (Sparrow, Balla, & Cicchetti, 1998)].

Standardized Tests

The Brazelton Neonatal Behavioral Assessment Scale (2nd ed.). The Brazelton Neonatal Behavioral Assessment Scale, 2nd ed. (NBAS-2, Brazelton, 1984), is a popular test of the neonate's current organizational and coping capacities in response to the stress of labor, delivery, and adjustment to the extra-utero environment. The NBAS-2 is designed for use with neonates 37 to 44 weeks of gestation age who do not require mechanical supports or oxygen. The test takes 20-30 minutes to administer (and about 15 minutes to record and score the infant's performance). Multiple assessments are recommended, with the first taking place at no earlier than three days of life. The NBAS data permit the construction of composite (factor and summative) scores, but the procedure is complicated and time-consuming. The NBAS inter-rater reliability estimates are quite high, but the indicators of the test-retest reliability are rather low, suggesting poor temporal stability for most items (Sameroff, 1978). The NBAS was originally designed for use with full-term healthy infants, but it has been used most extensively with premature and otherwise medically at-risk infants. The validity of the NBAS is supported by research that has

demonstrated its ability to discriminate groups of underweight, intra-utero drug or alcohol-exposed, gestationally diabetic, or intra-utero malnourished neonates as compared with the normative sample. The NBAS has been shown to predict infant-parent attachment and subsequent infant development, but primarily within the first year of life (Horowitz & Linn, 1982).

The Bayley Scales of Infant Development—II. The Bayley Scales of Infant Development-II (BSID; Bayley, 1993) represent the first restandardization of the Bayley test in 25 years. This scale is arguably the most widely used measure of the development of infants and toddlers. In addition, the BSID has an extensive psychometric history and a very respectable track record. The BSID-II is applicable to children from 1 through 42 months of age. The administration takes about 25 to 35 minutes for infants under 15 months of age and up to 60 minutes for children over 15 months.

The major difference between the older and the revised versions of the BSID is that the BSID-II is administered in "item sets," which are sets of items selected based on the age of the infant, whereas the original BSID used a continuous series of items. This modification created some confusion among infant assessors. For example, it is unclear which "item set" to use for infants born prematurely (Ross & Lawson, 1997a) or for infants living in cultural settings that differ from those of the infants in the normative sample (Gilliam, in press; Gilliam & Mayes, in press). For testers that adopt the corrected age procedure (for example, if a baby was born two months prematurely, and is 9 months old, it should be administered the set of items for 7-month-olds), the test developers recommend using the same "item set" that corresponds to the normative group used for determining that child's score. Specifically, if an infant's performance is to be compared to that of a typical 7-month-old, the examiner should administer the 7-month-old item set (Matula, Gyurke, & Aylward, 1997).

The BSID-II consists of three components: the Mental Development Index (MDI), the Psychomotor Development Index (PDI), and the Behavior Rating Scale (BRS). The MDI assesses the child's language development and problem-solving (cognitive) skills; the PDI assesses the child's gross and fine motor development; the BSR provides information on the child's behaviors during the assessment. The BSID-II permits obtaining age equivalence scores for four facets of development: Cognitive, Language, Social, and Motor.

The BSID-II was normed on 1700 infants representative of 1988 U.S. Census data. Test-retest reliabilities for time periods of 1 to 16 days range from .83 to .91 for the MDI and from .77 to .79 for the PDI. Stability for the BRS varies greatly depending on the age of the child, ranging from .55 to .90. Interrater reliability indicators for the BSID-II were reported to be .96 for the MDI, .75 for the PDI, and .70 for the total BRS. The total test internal-consistency reliability coefficients of the BSID-II are adequate, ranging from .89 (at ages of 2 1/2 years and 3 years) and .90 (at an age of 3 1/2 years). Concurrent validity of the MDI, as compared to other measures of general cognitive ability, typically falls in the .70 range, whereas the highest correlation between the PDI and other indicators of cognitive ability was .59. The norm tables for the BSID-II are adequate and are divided into one-month age blocks for children aged 36 to 42 months.

The Mental scale of the BSID-II has an adequate floor and good item gradients. This scale yields standard scores greater than 2 standard deviations below the mean for children between the ages of 2 1/2 years and 3 1/2 years. In addition, the scale has a number of items below the entry level for a child aged 2 1/2 years, providing adequate floors. The validity data on the BSID-II are still being accumulat-

ed, with the first data suggesting that validity of the instrument is adequate (e.g., Gerken, Eliason, & Arthur, 1994).

The qualitative characteristics of the BSID-II range from adequate to good. The manual contains information on the BSID-II theoretical framework, administration, scoring, and score interpretation. The instrument utilizes attractive and stimulating material and successfully alternates item types (e.g., language, visual, visual-motor). The administration time is about 60 minutes. The expressive language demands are minimal, but gestures are not acceptable responses. The test utilizes elements of dynamic testing, allowing for multiple trials and demonstrations. The BSID-II allows some flexibility in administration and includes rules for when to discontinue testing. The instructions are short and concise, even though the basic concept load (i.e., the number of basic concepts that are expected to be mastered by a child at a given age) is fairly high (28; Alfonso & Flanagan, 1999). The BSID-II has not yet been translated into languages other than English and at this time there are not yet norms from different cultures. Given that the BSID-II utilizes many objects from the traditional toy world of a North-American child, the amount of Western acculturation required to perform well on the test is relatively high.

The history of research with the BSID is replete with empirical demonstrations of both the usefulness and the futility of the data collected. On the one hand, the BSID has proven to be useful for the assessment of the current status of the infant (Lipsitt, 1992). On the other hand, the testing of children younger than 18 months of age with the BSID has yielded little predictive validity, at least to the extent that one is interested in anticipating the later intellectual or cognitive development of a given child (Colombo, 1993). As a matter of fact, Dr. Bayley herself expressed reservations about the use of the test for predictive purposes, suggesting that researchers examine the mother rather than the child. Researchers have arrived at the conclusion that, for children younger than 18 months, the BSID does not yield consistent results. In addition to lacking predictive power in the domain of intelligence, the Bayley does not predict either the child's behavioral scores or the child's psychiatric diagnoses (Dietz, Lavigne, Arend, & Rosenbaum, 1997). Burns et al. (1992) have made the case from their data that 3-month-olds are more like other 3-month-olds across a variety of tasks than they are like themselves over a long period of time. In other words, at 3 months of age, a normally-developing child and a mentally-retarded child appear to be very similar in their capacities as assessed by the BSID. When they reach the age of 3, however, they are very different. To sum up, the BSID appears to have some ability to predict which infants will score very poorly on intelligence tests before the age of 3, but shows a limited ability to accurately predict specific IQ scores, especially in average developing infants (Gibbs, 1990; Whatley, 1987).

The NEPSY. The NEPSY (NE from neuro and PSY from psychology, Korkman, Kirk, & Kemp, 1998) is a comprehensive battery designed to assess neuropsychological development in preschool and school-age children. The battery was designed to assess basic and complex aspects of cognitive development critical to children's ability to learn and be productive both inside and outside of school settings.

The NEPSY includes a set of neuropsychological subtests that can be used in various combinations. The assessment is carried out in five domains: Attention/Executive Functions, Language, Sensorimotor Functions, Visuospatial Processing, and Memory and Learning. The NEPSY's subtests are divided into core subtests and expanded subtests.

The theoretical background of the NEPSY is that of the Luria tradition of assessment (Christensen, 1984; Luria, 1973). According to this tradition, cognitive functions, such as attention and executive functions, language, movements, visuospatial abilities, and learning and memory, are complex capacities. They are composed of flexible and interactive subcomponents that are mediated by equally flexible, interactive neural networks. Therefore, it is important to identify and assess, as far as possible, both basic and complex subcomponents that contribute to performance within and across functional domains. The designers of the NEPSY capitalized on Luria's approach: Some subtests were designed to assess basic subcomponents of a complex capacity within single functional domains; other subtests were designed to assess subcomponents of cognitive functions that require contributions from several functional domains.

The NEPSY was standardized on a sample of 1,000 children (100 children in each of 10 age groups ranging in age from 3 through 12 years). The sample was representative of the U.S. general population on four indicators (gender, race/ethnicity, geographical region, and parent education level). All NEPSY subtest scores are z-transformed with a mean of 10 and standard deviation of 3.

Internal-consistency indicators and test-retest reliabilities range from inadequate (.42) to good (.91). Validity information on the NEPSY has been accumulated; initial reports (Korkman, Kirk, & Kemp, 1998) indicate the presence of only low-to-moderate correlations with core tests of cognitive functioning (e.g., WPPSI-R, WISC-III, and BSID-II).

Qualitative properties of the NEPSY are adequate. Test materials are engaging; tasks are interesting and stimulating. The NEPSY's manual is comprehensive and the instructions for administration are clear. To ensure that test-takers understand the instructions, the test utilizes examples and probes. The application of a discontinuation rule (i.e., a rule of interrupting a subtest after a certain number of failures) helps avoid frustration in children.

The NEPSY is one of the newest multi-domain batteries available on the market. It was initially created in Finnish. A parallel version was then developed in English. The battery is currently being scrutinized by a number of evaluators in different studies.

The Griffiths Mental Development Scales. The Griffiths Mental Development Scales are most popular outside of North America. The Griffiths consists of two tests: the Abilities of Babies (Griffiths, 1954), designed for infants from birth to 24 months, and the Abilities of Young Children (Griffiths, 1979), for children 24 months to 8 years old. The infant scale consists of five domains, modeled closely after the scales in Gesell's early work: Locomotor, Personal and Social, Hearing and Speech, Eye and Hand Coordination, and Performance. The test was normed on 571 infants from London, England. Reliability studies for the Griffiths have yielded mixed results, and validity studies have indicated relatively weak predictive ability for later IQ test scores (Thomas, 1970).

The Battelle Developmental Inventory. The Batelle Developmental Inventory (BDI, Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1984) assesses the development of children from birth through eight years of age. The BDI assessment time is longer than that for most similar tests, ranging, depending on the age of the child, from 1 to 2 hours. The BDI assesses development in the Personal-Social, Adaptive, Motor, Communication, and Cognitive domains. Each BDI domain is also divided into 5 subdomains

(to provide fine-grained information within each domain). When all domains are evaluated, the BDI produces a total developmental score.

The BDI was normed on a sample of 800 U.S. children. Several psychometric concerns were raised in a review of the BDI (McLinden, 1989). First, although the test authors reported exceptionally strong test-retest and inter-rater reliability for the BDI, a general lack of procedural details in the manual makes it difficult to evaluate these data adequately. Second, there is no information regarding the internal- consistency reliability of the BDI. Third, the concurrent validity studies were exceptionally small with respect to sample sizes, and the magnitudes of the correlations were rather low. Fourth, and most important, researchers have expressed major concerns about the BDI's normative data (Boyd, 1989). For the first two years, the BDI's normative data are presented in 6-month groups, whereas thereafter they are provided in 12-month groups. Therefore, a child's performance is compared to that of others who can be as many as 6 months older or younger for children under 24 months, or as many as 12 months older or younger for children older than 24 months. This lack of precision in the normative data tends to inflate standard scores for children who are old for their normative groups and to deflate standard scores for children who are young for their normative group (Boyd, Welge, Sexton, & Miller, 1989).

The Mullen Scales of Early Learning (*MSEL*; Mullen, 1995). The Mullen Scales of Early Learning are designed to assess children's development from birth to the age of 68 months. The Mullen takes about 15 to 60 minutes to administer, depending on the age of the child. The theoretical basis of the Mullen is a model of infant neurodevelopment, according to which the child should be assessed in five different domains: Gross Motor, Visual Reception (primarily visual discrimination and memory), Fine Motor, Receptive Language, and Expressive Language. The domain data can be combined into the overall Early Learning Composite score. The composite score based on the four cognitive scales (all but the Gross Motor Scale) represents so-called general intellectual ability.

Normative data for the Mullen are based on a sample of 1,849 children from the U.S.A. Internal-consistency reliabilities range from .75 to .83 for Mullen subtests and the reliability is .91 for the Early Learning Composite. Test-retest reliabilities range from .78 to .96, depending on the subtest. Inter-rater reliabilities range from .94 to .98. The indicators of concurrent validity are adequate.

The qualitative characteristics of the MSEL are adequate. The test contains many stimulating items that engage and maintain the interest of young children. The manual provides detailed descriptions of all manipulations, of scoring, and of how to interpret results.

The Peabody Picture Vocabulary Test-Revised (Dunn & Dunn, 1981). The Peabody Picture Vocabulary Test-Revised (PPVT-R) was originally developed in 1959 and then was revised in 1981. It is a nonverbal, multiple-choice test designed to evaluate the receptive vocabulary (assessed through hearing and by indicating "yes" or "no") of children and adults. The test is administered to individuals from age 2 1/2 years through adulthood. The PPVT-R requires no reading skill. The test is untimed. Testing time is 10-15 minutes.

The PPVT-R words were selected to be of equal numbers of nouns, gerunds, and modifiers in approximately 19 content categories. The words also were carefully selected to avoid gender, culture, religion, or race biases. The PPVT-R has two forms, L and M, with 175 plates in each form. Each plate contains four pictures. Items are arranged in increasing levels of difficulty. The two forms use different words in

pictures. The pictures are clearly drawn, and are free of fine details and interferences of background with the key stimulus or figure of the picture. Raw scores are converted into standard scores, with mean of 100 and standard deviation of 15.

Normative data were collected from a representative U.S. sample of 4,200 youths (aged 2 1/2 years through 18 years), and 828 adults (aged 19 through 40 years). Estimates of PPVT-R internal consistency, alternate-form reliabilities, and split-half reliabilities are somewhat lower than desirable, with most estimates between .75 and .85 (Slatter, 1992).

A number of studies have attempted to validate the PPVT-R. Specifically, the range of the PPVT-R/WISCR-R correlations is .16 to .86; the range of correlations between the PPVT-R and various measures of reading, language, and general achievement is .30 to .63 (for details, see Sattler, 1992). There is an ongoing discussion in the literature about what it is the PPVT-R measures: The test has been described as measuring language ability, verbal comprehension, vocabulary ability, receptive language, recognition vocabulary, verbal intelligence, vocabulary comprehension, vocabulary usage, comprehension of single words, single-word hearing vocabulary, single-word receptive vocabulary, and intelligence. Overall, the consensus has been reached that PPVT-R scores are not interchangeable with IQs.

A number of PPVT-R items have been found to be culturally biased against ethnic minorities (Argulewicz & Abel, 1984; Reynolds, Willson, & Chatman, 1984). Therefore, the PPVT-R should be applied with caution in populations other than North American native speakers of English.

Behavior Rating Scales

Behavior rating scales typically require evaluation of a number of specific behaviors or correlates of behavior organized into empirically-derived factors or scales. These scales are expected to provide information about a child across social-emotional or adaptive behavior dimensions. Typically, such scales have originated from literature reviews and/or the assessment of behavioral peculiarities of children who are representative of the target population of the rating scale. These descriptors are then factor-analyzed and the resulting scales are evaluated for their psychometric soundness (i.e., reliability and validity).

A typical administration of a behavioral rating scale includes respondents' (usually parents' and teachers') ratings of the degree to which certain behavioral descriptors are present in the child's behavior (e.g., "Does your child tease other children? /often, sometimes, never"). The responses are summed up across the behavioral factors and then compared to some standardization sample or reference group.

Child Behavior Checklist. The Child Behavior Checklist exists in five different forms: the Child Behavior Checklist (ages 2 to 3), the Child Behavior Checklist (ages 4 to 18), the Teacher's Report Form (ages 5 to 18), the Youth Self-Report (ages 11 to 18), and the Direct Observation Form (ages 5 to 14). Of these instruments, only the first two forms, completed by parents, can be used with preschoolers.

The Child Behavior Checklist for ages 2 to 3 (CBCL/2-3; Achenbach, 1991) consists of 99 items and 1 open-ended item describing various behaviors, emotional problems, or reactions to specific situations. The CBCL items are rated on a 3-point scale (2—Very true or often true; 1—Somewhat true or sometimes true; and 0—Not true at the present time or over the last two months). The CBCL/2-3 has three

global scales (Total Problems, Internalizing, and Externalizing) and six narrow-band scales (Social Withdrawal, Depressed, Sleep Problems, Somatic Problems, Aggressive, and Destructive).

The Child Behavior Checklist for ages 4 to 18 (CBCL/4-18; Achenbach, 1991) is a standardized parent-report measure of children's adaptive competencies and problem behaviors that is widely used internationally in clinical and research settings. The measure consists of 20 competence items and 118 items describing behavioral/emotional problems. The total Competence scale includes scores on the scales of Activities, Social, and School. The problem behaviors are scored on eight factor-based, narrow-band scales: Withdrawn, Somatic Complaints, Anxious/Depressed, Social Problems, Thought Problems, Attention Problems, Delinquent Behavior, and Aggressive Behavior. In addition, two broad-band scales can also be scored: Internalizing and Externalizing. Finally, there is also a Total Problem Score that provides an overall index of the number and severity of reported problem behaviors.

The CBCL manual (Achenbach, 1991) contains details regarding standardization, factor analyses, and other psychometric properties. The CBCL/2-3 and CBCL/4-18 appear to be psychometrically strong. The scale statements are written at the fifth-grade reading level. The completion of the scales usually takes about 20 minutes. The CBCL scales are important instruments to be used in a preschool assessment battery. They evaluate children's competencies and problem behaviors and provide additional information that cannot be obtained through cognitive testing.

The Conners' Behavior Rating Scales. The Conners' Behavior Rating Scales consist of the Conners' Parent Rating Scales (CPRS) and the Conners' Teacher Rating Scales (CTRS). These scales were originally developed to identify hyperactive children but have since been expanded to identify children with other, related behavioral problems. Each scale is available in two forms, short and long. The CTRS items are rated along a 0 to 3 scale (0 = Not at All; 1 = Just a Little; 2 = Pretty Much, 3 = Very Much). The parent-version items are structured into five scales: Conduct Problem, Learning Problem, Psychosomatic, Impulsive-Hyperactive, and Anxiety. The teacher-version items form three scales: Conduct Problem, Hyperactivity, and Inattentive-Passive. Psychometric properties of the CPRS range, for different scales, from poor to adequate.

The Vineland Adaptive Behavior Scales. The Vineland Adaptive Behavior Scales (VABS; Sparrow, Balla, & Cicchetti, 1984) exist in three versions (the Survey Form, the Expanded Form, and the Classroom Edition). This instrument is used to assess the ability of handicapped and nonhandicapped children to perform the daily activities required for personal and social sufficiency from the ages of birth through 19 years. The VABS (all three versions) measure adaptive behavior using four specific domains: Communication, Daily Living Skills, Socialization, and Motor Skills (with this last scale administered only to children 0-6 years of age). Each of the four primary VABS domains contains specific subdomains: The Communication domain is divided into receptive, expressive, and written subdomains; the Daily Living Skills domain is divided into personal, domestic, and community subdomains; the Socialization domain is divided into interpersonal relationships, play and leisure time, and coping skills subdomains; and the Motor Skills domain is divided into gross and fine motor subdomains. The subscale scores are added up to yield an Adaptive Behavior Composite. The Survey and Expanded Forms of the VABS have an optional Maladaptive Behavior subscale. The Survey Form consists of 297 items and takes up to 60 minutes to administer; the Expanded Form consists of 577 items and takes about 90 minutes to administer. Both forms are administered as semi-structured interviews with a par-

ent or significant caregiver. The Classroom Edition consists of 244 items and takes about 20 minutes to complete. This edition is completed by a teacher.

For the Survey and Expanded Forms, the standardization was carried out on a 3,000-individual U.S. sample, representative of the general population by gender, race/ethnicity, community size, geographical region, and SES. A similarly stratified sample of 3,000 students was used for the Classroom Edition. The median split-half reliabilities of the VABS range from .83 to .95; test-retest reliabilities are in the .80s and .90s; inter-rater reliabilities range from .62 to .75. Validity indicators have been found to be satisfactory (e.g., Atkinson, Bevc, Dickens, & Blackwell, 1992). However, it appears that, for preschoolers, some of the domains may have inadequate floors (Knoff, Stollar, Johnson, & Chenneville, 1999).

Theories-of-Cognition-Based Tests

The majority of intelligence tests have been developed within the psychometric paradigm, an approach based on the identification of abilities (verbal and spatial abilities, memory, reasoning, etc.) through the factor analysis of sets of diverse cognitive tasks. Most modern psychometric tests (but not all, as shown below) address both a general factor (the so-called *g*-factor, reflecting the positive manifold of correlations between various cognitive abilities) and distinct, though correlated, group factors. Whereas all of the psychometric tests have a full-scale or a composite index that, presumably, reflects the *g*-factor, no single test completely overlaps with any other test in terms of all the cognitive abilities that are measured.

McGrew and Flanagan (1996) conducted a review of tests of intelligence for young children and classified the subtests of all major intelligence batteries according to the Horn-Cattell G_f-G_c theory (Horn, 1991, 1994; Horn & Noll, 1997) and the Three-Stratum Theory of Cognitive Abilities (Carroll, 1993, 1997). Presumably, such a classification provides a unified theoretical scheme for comparing and contrasting different tests of intelligence (Alfonso & Flanagan, 1999).

The Wechsler Preschool and Primary Scale of Intelligence—Revised. The Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R; Wechsler, 1989) is the most recent version of a test that was initially developed in the late 1960's (Wechsler, 1967). The test is an individually administered clinical instrument for assessing the intelligence of children aged 3 years through 7 years, 3 months. The WPPSI-R is organized so that one group of subtests (Information, Comprehension, Arithmetic, Vocabulary, Similarities, and Sentences) yields a Verbal IQ and another group of subtests (Object Assembly, Geometric Design, Block Design, Mazes, Picture Completion, and Animal Pegs) yields a Performance IQ. The Verbal and Performance IQs combine to yield the Full Scale IQ, which is interpreted as a measure of so-called general intellectual functioning. Descriptions of all WPPSI-R subtests can be found in Gyurke (1991) and in Wechsler (1989).

The WPPSI-R subtests have a mean of 10 and a standard deviation of 3. The Verbal, Performance, and Full scales have a mean of 100 and a standard deviation of 15. Administration time for the WPPSI-R is usually somewhat greater than one hour.

Overall, the WPPSI-R displays adequate standardization characteristics, with a total standardization sample of 1,700 individuals, 400 per each one-year interval. The sample is representative of the general U.S.

population with respect to gender, race/ethnicity, geographic location, and SES. The WPPSI-R norms were collected in 1984-1985; thus, current WPPSI-R scores may slightly overestimate cognitive ability.

The total internal-consistency reliability coefficients of the WPPSI-R are .95 or greater. The WPPSI-R test-retest reliability is estimated at .91, but the procedures used to obtain it are considered inadequate, given that the test-retest reliability sample comprised children past the preschool age.

The WPPSI-R's floors are inadequate at the early age levels of the test (i.e., 2 years and 11 months of age and slightly above). All subtests have adequate floors only by the age of 4 years and 9 months. The subtest item gradients are generally adequate.

The WPPSI-R has adequate validity. Researchers conducted several factor-analytic studies on the WPPSI-R data (e.g., Allen & Thorndike, 1995; Stone, Gridley, & Gyurke, 1991) and consistently obtained a two-factor solution, supporting the claim that a Verbal-Performance dichotomy underlies the WPPSI-R.

There is much validity data available. Dozens of studies have been conducted that compare the WPPSI-R with other major intelligence tests, such as the Fourth Edition of the Stanford-Binet—SB-IV (median $r = .78$, Thorndike et al., 1986b), the McCarthy Scales of Children's Abilities ($r = .86$, Sattler, 1992), the Woodcock-Johnson Revised—WJ-R (e.g., $r = .70$, Harrington, Kimbrell, & Dai, 1992), and the Differential Aptitude Scales—DAS (e.g., $r = .74$ with DAS; Elliott, 1990b). In general, correlations between total scores on the WPPSI-R and scores on other instruments have been moderate to high (McGrew & Flanagan, 1996).

The qualitative characteristics of the WPPSI-R range from poor to adequate (Table 2). The WPPSI-R Manual (Wechsler, 1989) provides adequate information about the development of the instrument, its underlying constructs, administration and scoring procedures, and interpretations. The test materials are engaging and are likely to attract the attention of a preschool child; moreover, some subtests include stimulating initial tasks (i.e., Object Assembly). The WPPSI-R nicely alternates verbal and nonverbal subtests. Moreover, this test successfully utilizes elements of dynamic testing by including teaching items, second trials, and demonstrations by the examiner.

However, the WPPSI-R has a number of drawbacks. The major one is the length of the test: Many young children cannot remain focused and attentive during the entire administration of the test. The WPPSI-R does not provide alternative stopping rules, which makes the test rather frustrating for young children. In addition, many WPPSI-R subtests rely heavily on extensive expressive language skills (e.g., Comprehension and Vocabulary). Moreover, the WPPSI-R does not build in gestures as possible answers. In addition, the WPPSI-R directions are complex and unnecessarily include many basic concepts (up to 42, according to Flanagan et al., 1995).

There are no core directions on how to translate and adapt the test to cultures other than the North American one, and there are no norms for children from other cultures. As for the degree of the importance of North-American acculturation necessary for successful performance on the test, it is assumed to be high for the subtests constituting the Verbal scale, and moderate-to-low for the subtests constituting the Performance scale. Thus, the WPPSI-R is of limited utility in the evaluation of children from diverse cultural backgrounds.

Overall, the WPPSI-R possesses strong psychometric characteristics; in many respects, it is quantitatively the strongest instrument in the field of early child assessment (Kamphaus, 1993). Yet, the test is too long, its administration is too complex, it does not adequately handle failure-related frustration, the acceptability of answers is heavily dependent on the child's expressive language, and it is over-reliant on the values of North-American culture.

The Wechsler Intelligence Scale for Children—Third Edition. The Wechsler Intelligence Scale for Children—Third Edition (WISC-III; Wechsler, 1991) is the most current edition of a test that was initially developed in the late 1940s (Wechsler, 1949). This test is an individually administered clinical instrument for assessing the intellectual abilities of children aged 6 years through 16 years, 11 months. The instrument consists of three main composite scores: Verbal IQ (comprising Information, Similarities, Arithmetic, Vocabulary, Comprehension, and Digit Span subtests), Performance IQ (comprising Picture Completion, Coding, Picture Arrangement, Block Design, Object Assembly, Symbol Search, and Mazes subtests), and Full Scale IQ. Although based originally on a conception of intelligence that emphasized the pervasive nature of so-called general intelligence, the current edition of the WISC offers scores for four factors (Verbal Comprehension, Perceptual Organization, Processing Speed, and Freedom From Distractibility).

The Stanford-Binet Intelligence Scale: Fourth Edition. The Stanford-Binet Intelligence Scale: Fourth Edition (SBIS-IV, Thorndike, Hagen, & Sattler, 1986a, 1986b) is an individually administered intelligence test used to assess the cognitive abilities of individuals from age 2 years to adult. The Fourth Edition is the latest version of the Stanford-Binet, which was originally published in 1916. The SBIS-IV is based on a three-level hierarchical model consisting of "g" (a general reasoning factor) and three second-order factors (Crystallized Abilities such as Verbal Reasoning and Quantitative Reasoning, Abstract/Visual Reasoning, and Short-Term Memory). The Verbal Reasoning area score is derived from the Vocabulary, Comprehension, and Absurdities subtests; the Abstract/Visual Reasoning score is derived from the Pattern Analysis and Copying subtests; the Quantitative area score is derived from the Quantitative subtest; and the Short-Term memory score is based on the Bead Memory and Memory for Sentences subscores (for subtest descriptions, see Delaney & Hopkins, 1987; Glutting & Kaplan, 1990). Scores from one or more of the SBIS areas are combined to yield the Test Composite (a measure of *g*).

SBIS subtest scores have a mean of 50 and a standard deviation of 8. Area scores and the Composite scores have a mean of 100 and a standard deviation of 16. The SB-IV was standardized on 5000 individuals with at least 200 individuals per one-year interval. The normalization sample was, as a whole, representative of the U.S. population (in terms of gender, geographic region, race/ethnicity, and community size), with somewhat under-represented low SES and over-represented high SES participants; no data on age-specific representativeness are available.

The total test internal-consistency reliability coefficients are good, with the lowest coefficient being .95. The test-retest reliabilities are also good (.91), but may be biased because the sample suffers from non-representativeness. The most significant psychometric limitations of the SBIS-IV are subtest and area floors and item gradients. All subtests (and, correspondingly, all areas) appear to have inadequate floors for 0-2-year-olds and the floors become adequate only at about age 5; thus, the SB-IV is inadequate for the assessment of very young children. The item gradients are inadequate for six of the eight subtests for preschoolers (i.e., Comprehension, Absurdities, Bead Memory, Quantitative, Copying, and Pattern

Analysis) at ages 2.6 to 3.5; thus, fine gradations in ability may not be detected on most subtests of the SB-IV, especially in the case of very young children (Alfonso & Flanagan, 1999).

Overall, the validity of the SB-IV is considered adequate. However, the quality of the construct-validity evidence has been questioned (Glutting & Kaplan, 1990; Kaplan & Alfonso, 1997). Specifically, although there are convergent data suggesting that the Test Composite of the SB-IV can be interpreted as a measure of general intelligence, there is considerable controversy regarding the factor structure of the instrument (in terms of both factor number and factor loadings). Out of many studies attempting to investigate the factor structure of the SB-IV, not a single one has supported the test authors' claim of its invariance across age (e.g., Keith, Coll, Novak, White, & Pottebaum, 1988; Kline, 1989; Molfese, Yaple, Helwig, Harris, & Connell, 1992). Despite the lack of agreement regarding the construct validity of the SB-IV across the age range of the test, however, there is general consensus regarding the two-factor structure of the SB-IV for preschoolers (e.g., Molfese et al., 1992; Ownby & Carmin, 1988), with Vocabulary, Comprehension, Absurdities, and Memory for Sentences forming a Verbal Comprehension factor and Pattern Analysis, Copying, Quantitative, and Bead Memory forming a Nonverbal Reasoning/ Visualization factor (Sattler, 1992). As for criterion validity, dozens of studies have been carried out comparing the SB-IV with other major intelligence tests, such as the WPPSI-R (e.g., Carvajal et al., 1988; McCrowell & Nagle, 1994), WISC-R (e.g., Brown & Morgan, 1991; Phelps, Bell, & Scott, 1988), K-ABC (e. g., Lamp & Krohn, 1990; Rothlisberg, & McIntosh, 1991), and others (for more detail, see Appendix). In general, correlations between the total scores of the SB-IV and other instruments have been moderate to high (McGrew & Flanagan, 1996) and test scores for exceptional groups (e.g., learning disabled, gifted) have been essentially similar (Sattler, 1992). Therefore, the SB-IV appears to be a valid measure of many aspects of intellectual functioning for children of most ages and for a variety of exceptional subpopulations.

The qualitative characteristics of SB-IV are adequate. The theoretical foundation of the test and its psychometric properties are presented in the SB-IV kit's manuals (Thorndike et al., 1986). Guidelines for interpreting individual results and specific details of administration are provided in a stand-alone handbook (Delaney & Hopkins, 1987). The SB-IV contains attractive manipulative materials, but because they are downward extensions of test items for older children, they could use some modification that would result in a higher degree of engagement of younger children.

The SB-IV is effective in alternating verbal and nonverbal subtests. The completion time, on average, is an hour. However, the administration of some subtests is rather awkward, sometimes requiring change of stimulus material or shift in instructions. The SB-IV requires minimum expressive language; gestures are not acceptable responses. The instructions are lengthy and use a large number of basic concepts (up to 25, Alfonso & Flanagan, 1999). Only four of the eight subtests of the SB-IV for preschoolers utilize elements of a dynamic-testing approach by including sample items and providing young children an opportunity to learn the task. In addition, the test does not have alternative stopping rules (e.g., the option of stopping the test after an established number of consecutive failures).

The Verbal Reasoning cluster subtests are highly sensitive to North American acculturation. The Short-Term Memory subtest is moderately sensitive, and the Abstract/Visual Reasoning subtests are only slightly sensitive.

The test directions of the SB-IV and children's verbal responses are not commercially available in languages other than English. There are no norms for individuals from cultures outside the U. S. A.

Overall, the SB-IV is characterized by adequate psychometric indicators. It is considered to be a valid measure of so-called *general* intellectual functioning. The data on construct validity are somewhat inconsistent and the factor structure of the instrument is an unresolved question. The test is well-described (i.e., its theoretical framework is clear and its structure is transparent), has minimal expressive language requirements, and takes about 60 minute to administer. But the instructions are wordy and the administration of the test is sometimes cumbersome. Two main weaknesses of the SB-IV are the lack of interpretability of area-specific subtests and inadequate floors and item gradients.

The Cattell Infant Intelligence Scale. The Cattell Infant Intelligence Scale (Cattell, 1960) was conceptualized as a downward extension of the 1937 Stanford-Binet Intelligence Scale. The Cattell was designed to assess infants and toddlers from 2- through 30-months-old. In constructing the instrument, Cattell relied heavy on Gesell's work, using the same or similar items as did Gesell in his instrument. In order closely to match the *g*-factor paradigm, however, items addressing gross motor and personal-social development were excluded. The reliability indicators were adequate, but the predictive validity was low (the correlations between the Cattell and the SBIS at 36 months were very low for 2-year-olds, but somewhat higher for 24-30-months-olds; Thomas, 1970). Overall, though designed specifically as an extension of the *g*-based tests to early childhood, the Cattell appears unable to improve the predictive power of indicators of cognitive development in early childhood.

The McCarthy Scales of Children's Abilities. The McCarthy Scales of Children's Abilities (McCarthy, 1972) form a well standardized and psychometrically sound measure of the cognitive abilities of young children (ages 2 1/2 to 8 1/2 years). The test is individually administered and takes about 45 to 60 minutes to administer, depending on the age of the child. The McCarthy Scales have some unique features valuable for the assessment of young children with learning problems or other exceptionalities (Sattler, 1992). The test produces a general measure of intellectual functioning called the General Cognitive Index (GCI), as well as a profile of abilities that includes measures of verbal ability, nonverbal reasoning ability, number aptitude, short-term memory, coordination, and hand dominance.

The scale indices derived from the McCarthy Scales subtest are standard scores, with a mean of 50 and a standard deviation of 10. The overall General Cognitive Index has a mean of 100 and a standard deviation of 16. This index is considered to be an indicator of the child's ability to integrate his or her accumulated knowledge and to adapt that knowledge in order to perform the tasks on the scales.

The standardization of the McCarthy Scales was excellent ($N = 1032$). The sample was representative of the general population of the U.S.A. (on the variables of age, sex, race and ethnicity, geographic region, father's occupation, and urban-rural residence).

The psychometric properties of this scale are also very good, with median split-half, internal-consistency, and test-retest reliabilities ranging between .85 and .93 (Sattler, 1992). The concurrent validity of the McCarthy Scales is acceptable with the Stanford-Binet, WISC, WISC-R, WPPSI, and K-ABC (with correlations ranging from .45 to .90). Construct validity, however, appears to be questionable, with different numbers of factors revealed in different studies and for boys and girls.

The McCarthy Scales' Manual is comprehensive and easy to use. Materials are well-constructed and appeal to children.

The Differential Ability Scales. The Differential Ability Scales (DAS; Elliot, 1990a) form an individually administered battery of cognitive and achievement tests for children and adolescents from ages 2 1/2 years through 17 years. The Cognitive Battery is organized into a set of core subtests that yield a General Conceptual Ability (GCA) score and a set of diagnostic subtests that provide additional information on specific abilities. There is also an intermediate layer of so-called cluster scores, linking specific subtests to the GCA score. The structure of the test is flexible and age-dependent. Thus, for children aged 2 years, 6 months to 3 years, 5 months, there are no cluster scores because abilities are relatively undifferentiated at this young age. The GCA score is a function of scores on core subtests (Block Building, Verbal Comprehension, Picture Similarities, and Naming Vocabulary) and diagnostic subtests (Recall of Digits, Recognition of Pictures). For children aged 3 years, 6 months, to 5 years, 11 months, the ability clusters are verbal (core subtests are Verbal Comprehension and Naming Vocabulary) and nonverbal (core subtests are Picture Similarities, Pattern Constructions, and Copying), and the diagnostic subtests are Block Building, Matching Letter-Like Forms, Recall of Digits, and Recognition of Pictures. For children aged 6 years to 17 years 11 months, the clusters are verbal ability (including core subtests of Word Definition and Similarities), nonverbal reasoning ability (core subtests of Matrices, and Sequential and Quantitative Reasoning), and spatial ability (including subtests of Recall Designs and Pattern Construction). The diagnostic subtests are Recall of Digits, Recall of Objects, and Speed of Information Processing. The GCA is viewed as providing an estimate of so-called general intelligence (Elliott, 1990b).

Elliot does not provide any explicit theoretical framework underlying the DAS. McGrew and Flanagan (1996), however, view this instrument as yet another realization of the G_f-G_c theory.

All DAS subtests have a mean of 50 and a standard deviation of 10, whereas all composites have a mean of 100 and a standard deviation of 15. The administration time ranges between 35 and 60 minutes; the time is determined by whether the diagnostic subtests are administered (Elliott, 1990a).

The DAS was standardized on a U.S.-population sample, representative of the general U.S. population (by gender, geographic region, race/ethnicity, enrollment in educational programs, and SES). The sample included 3,475 individuals, with 200 to 350 individuals per one-year interval. The DAS norm tables are saturated in blocks of 3 months. The DAS was standardized in 1987-1988; therefore, these norms are still adequate, although barely so.

The DAS demonstrates adequate reliability and validity. Internal-consistency reliability coefficients are .90 or higher across the preschool range, with somewhat lower coefficients (.89) for children between the ages of 0-3 and 4-6 years. The test-retest evaluation was carried out on a representative sample, with an interval of four weeks, and was found to be .90.

All DAS subtests but Verbal Comprehension have demonstrated adequate floors; the Verbal Comprehension subtest floor becomes adequate by the age of 4 years and 4 months. However, all DAS composites have adequate floors across the preschool age range. The DAS generally has adequate item gradients at the middle and upper end of the preschool range, but two subtests (Block Design and Naming Vocabulary) have inadequate item gradients at the lower preschool age.

Factor analyses (e.g., Keith, 1990) suggest a one-factor (*g*-factor) solution for the DAS subtest data collected on the youngest children, and a two-factor solution for the data collected on older preschoolers (starting at the age of 3 years and 6 months). As for criterion validity, dozens of studies have been carried out comparing the DAS with other major intelligence tests, such as the WPPSI-R, the Woodcock-Johnson (WJ-R) , and Kaufman Assessment Battery for Children (K-ABC). In general, correlations between the total scores on the DAS and other instruments have been moderate to high (McGrew & Flanagan, 1996). Therefore, the DAS appears to be a valid measure of intellectual functioning for most ages and for a variety of exceptional subpopulations.

The qualitative characteristics of the DAS range from adequate to good (Table 2). The manual contains sufficient information on the theoretical background, administration, scoring, interpretation, and psychometric properties of the test. The DAS materials are age-appropriate, stimulating, and engaging. To involve children in the process of testing, the DAS alternates verbal and nonverbal subtests, beginning with stimulating tasks. Many of the DAS subtests require minimum expressive language skills, and gestures are acceptable responses. The DAS utilizes elements of dynamic testing by including sample and teaching items, second trials, and demonstrations to ensure that the child has understood the task. Another positive characteristic of the DAS is the inclusion of stopping rules. The main limitation of the test is the length of direction and the high number of basic concepts (23; Alfonso & Flanagan, 1999).

The DAS has not been translated into languages other than English and does not include a system of core assumptions for translations. There are no norms available for individuals from other than North-American cultures.

Similar to other *g*-based tests of intelligence, the importance of North-American acculturation for the DAS subtests ranges from high to low. The Verbal subtests (i.e., Verbal Comprehension and Naming Vocabulary) are highly dependent on exposure to North American culture, whereas the subtests comprising the Nonverbal cluster (i.e., Copying, Pattern Construction, and Block Building) and the Special Nonverbal Composite (Block Building and Picture Similarities) are perhaps somewhat less susceptible to the impact of culture.

Overall, the DAS, more than any other *g*-based instrument, appears to achieve a balance between good quantitative and qualitative indicators and, therefore, is highly regarded by professionals in the field of early childhood assessment (Alfonso & Flanagan, 1999).

The Woodcock-Johnson Psycho-Educational Battery-Revised: Tests of Cognitive Ability. The Woodcock-Johnson (*WJ-R COG,* Woodcock & Johnson, 1989) is designed for individuals aged 24 months through 95+ years. The battery contains 21 tests of cognitive ability divided into standard and supplemental batteries. The test is specifically based on the Horn-Cattell G_f-G_c theory (Horn, 1991, 1994; Horn & Noll, 1997) and the Three-Stratum Theory of Cognitive Abilities (Carroll, 1993, 1997), so the standard battery contains seven tests, one measure for each of seven G_f-G_c factors (Crystallized Ability, Short-Term Memory, Visual Processing, Auditory Processing, and Long-Term Retrieval). The supplemental battery contains 14 tests. Of these, the first 7 tests (i.e., tests 8 through 14) provide complementary measures of the seven G_f-G_c Cognitive Ability Clusters. The remaining seven tests on the WJ-R COG supplemental battery (tests 15-21) provide mixed measures of G_f-G_c abilities and may be administered to derive additional information about an individual's cognitive strengths and weaknesses. Of the WJ-R COG's 21 subtests, only five [Picture Vocabulary (a test of Crystallized Ability), Memory

for Sentences (a test of Short-Term Memory), Visual Closure (a test of Visual Processing), Incomplete Words (a test of Auditory Processing), and Memory for Names (a test of Long-Term Retrieval)] are applicable to children between the ages of 2 years and 5 years, 11 months. Combined, scores on these subtests yield the Broad Cognitive Ability Early Development (BCA-ED) cluster, which is interpreted as an estimate of general intellectual functioning (Woodcock & Mather, 1989).

The WJ-R COG subtests have a mean of 100 and a standard deviation of 15. The time required for administration of the BCA-ED ranges from 20 to 30 minutes.

The WJ-R COG standardization was carried out on a representative sample (approximating the U. S. population on gender, geographic region, race/ethnicity, community size, and SES) comprising 6,359 individuals with at least 100 individuals per one-year interval. However, only 705 children were included in the preschool sample, and no specific age-based information was reported on this sample. The norm table of the WJ-R COG is divided into one-month age blocks for children between the ages of 2 and 5 years, 11 months.

The WJ-R internal consistency estimates are good ($\geq .93$ across the preschool age range). The test-retest reliability sample included very few preschoolers and, therefore, is considered inadequate. However, the overall test-retest reliability indicator for the WJ-R COG for this sample is .87. All BCA-ED subtests (except one: Incomplete Words) have adequate floors and item gradients throughout all preschool years. Therefore, the BCA-ED not only provides estimates of ability in the borderline range (i.e., at least 2 standard deviations below the normative mean) for very young children, in the middle and upper end of the preschool age range, but it is also sensitive to fine gradations in ability between the mean and 1 to 2 standard deviations below the mean (Alfonso & Flanagan, 1999).

There is ample evidence supporting the construct validity of WJ-R COG, but most of this evidence has been collected for individuals aged 5 to 80+. Moreover, the BCD-ED Cluster has only one indicator (subtest) per ability. Therefore, Alfonso and Flanagan (1999) suggest interpreting the battery for preschoolers as a measure of general cognitive ability (so-called general intelligence). The criterion validity indicators are good, demonstrating a moderate to high degree of similarity among intelligence tests for preschoolers (McGrew & Flanagan, 1996).

The qualitative characteristics of WJ-R COG range from poor to adequate. The manual is comprehensive, containing the necessary information on the test's underlying theory, development, psychometric properties, administration, scoring, and interpretation. However, test material does not include manipulatives, and appears to be of rather low interest to children. Moreover, the tests' answers emphasize expressive language and do not accept gestures as adequate responses.

The WJ-R COG is the only cognitive test for young children that has a parallel form in Spanish (Batería-R COG, Woodcock & Muñoz-Sandoval, 1996). The instructions are brief and comprehensive, with a minimum number of basic concepts (12, Alfonso & Flanagan, 1999). The WJ-R COG does not utilize elements of dynamic testing. Moreover, the test does not include alternative stopping rules.

Like the other tests discussed here, the WJ-R COG's dependency on North American acculturation is distributed unevenly (Hessler, 1993); this dependency is quite pronounced for some subtests (e.g., Picture Vocabulary) and less so for others (e.g., Memory for Names).

The WJ-R COG is unique in that it has subtest floors and item gradients that are generally good throughout the preschool age range, rendering it a sensitive instrument for quantifying fine gradations in ability across various levels of cognitive functioning. In general, the WJ-R COG is designed to measure a broader range of abilities than is usually captured by intelligence tests. Unfortunately, these abilities are underrepresented in the WJ-R COG variant for young children. Moreover, due primarily to the limited attractiveness of the stimuli for younger children, the lack of discontinuation rules, and a certain dependency on verbal responses, the test is rarely used in early-child assessment.

The Kaufman Assessment Battery for Children. The Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983) measures both intelligence and achievement. It is designed to assess functioning in both normal and exceptional children of ages 2 1/2 through 12 1/2 years. Four global indicators of functioning are assessed: Sequential Processing, Simultaneous Processing, (combined into the Mental Processing Composite), Nonverbal Performance, and Achievement (an indicator of the optimality of the application of the two types of processing in the context of academic-skills mastery). There are a total of 16 subtests (3 sequential, 7 simultaneous, and 6 achievement), but not all subtests are administered at every age (no more than 13 are administered to any one child). Only three subtests (Hand Movements, Gestalt Closure, and Faces and Places) run throughout the ages covered by the battery. The K-ABC is intended for use in school and clinical settings, with an administration time being approximately 45 minutes for preschool children and about 75 minutes for those of school-age children.

Unlike the other tests, this test has a solid theoretical basis and draws on Alexander Luria's (1980) theory of the functional systems of the brain. The theoretical framework underlying the K-ABC makes a distinction between sequential and simultaneous mental processes. Sequential processing refers to the child's ability to solve problems by mentally arranging input in sequential or serial order. This type of processing is crucial for such mental operations as learning grammatical relationships and rules, understanding the chronology of events, and making associations between sounds and letters. Correspondingly, the Sequential Subscale contains three subtests: Hand Movements, Number Recall, and Word Order. Simultaneous processing refers to the child's ability to synthesize information in order to solve a problem. This type of processing is fundamental to learning the shapes of letters, deriving meaning from pictorial stimuli, or determining the main idea of a story. The Simultaneous Scale contains seven subtests: Magic Window, Face Recognition, Gestalt Closure, Triangles, Matrix Analogies, Spatial Memory, and Photo Series. The Achievement Scale contains six subtests: Expressive Vocabulary, Faces and Places, Arithmetic, Riddles, Reading/Decoding, and Reading/Understanding. The score on Nonverbal Performance is composed of the scores of those subtests that form the Sequential and Simultaneous Processing Scales (Face Recognition, Hand Movement, Triangles, Matrix Analogies, Spatial Memory, and Photo Series) that do not require words.

The Sequential and Simultaneous Processing Scales were designed to reduce the effects of verbal processing. Moreover, the test was intentionally designed to minimize the effects of gender, ethnic, and North-American cultural bias.

Raw scores for the subscales are converted into scaled scores with a mean of 10 and a standard deviation of 3; the global scales are transformed into standard scores with a mean of 100 and a standard deviation of 15. The standardization of the K-ABC was adequate; it was conducted on a large North-American sample ($N = 2000$). However, the sample shows some representation problems with Hispan-

ic-Americans. In addition, low-educational-level Blacks were significantly underrepresented. Internal consistency reliabilities are satisfactory, ranging from .86 to .97 for composite scales. Stability of the K-ABC as assessed by the means of test-retest reliability is also adequate (Sattler, 1992). The K-ABC was validated against many other tests of children's cognitive functioning (see Appendix). The concurrent-validity indicators are satisfactory. Item distribution characteristics are adequate, but the K-ABC has a low ceiling that may limit its usefulness in evaluating gifted children. Over half of the subtests on the Simultaneous and Sequential Processing Scales provide maximum scores that are only 2 standard deviations or less above the mean. The Achievement Scale also has a restricted range.

The K-ABC is recognized as an instrument useful in certain situations, especially those requiring emphasis on nonverbal cognitive abilities. However, the K-ABC is not recommended for use as the primary instrument for identifying the intellectual abilities of normal or special children either in research or in clinical settings (Sattler, 1982).

The Standard Raven Progressive Matrices (SPM; Raven, 1960), drawing on Spearman's (1923, 1927) theory of general ability, consists of 60 matrix problems, which are separated into five sets of 12 designs each. Within each set of 12, the problems become increasingly difficult. Each individual design has a missing piece. The participant's task is to select the correct piece to complete the design from among six to eight alternatives. Correct responses are based on various organizing principles, such as increasing size, reduced or increased complexity, and number of elements. The SPM uses nonverbal stimuli, and it is assumed that it does not require a specific knowledge base. A separate test, referred to as Coloured Progressive Matrices (Raven, 1965), has been developed for children in the 5-11 age range and the elderly (65+ years of age). Similarly, persons believed to be of high intellectual ability can be administered the Advanced Progressive Matrices (Raven et al., 1992). The SPM is considered one of the most reliable instruments for measuring general intelligence, especially its fluid aspects (Court, 1988; Raven, 1989). The latest edition of the tests was published in 1995. This test series generally is not appropriate for preschoolers.

Criterion/Norm-Referenced Assessment

Norm-referenced comparison is the most commonly used method of comparison. It is especially prevalent in mental-health assessment and involves comparing particular observed behaviors to those of a large representative sample of children. In other words, diagnostic and screening tests are used to compare a child's current level of functioning to that of other children. *Criterion-based comparison* usually involves comparing a child's performance to some set of expectations or set of standards, such as those indicative of school readiness. Criterion-based instruments typically include many different items intended to reflect all important developmental stages and competencies of various ages.

The Brigance Diagnostics Inventory of Early Development—Revised. The Brigance (*BDIED,* Brigance, 1991) is one of the most popular criterion-referenced tests used with young children (from birth to 7 years). The Brigance surveys skills in 12 different developmental domains, including social and emotional, communicative, motor, and pre-academic skills (reading, math, and manuscript writing). There is, however, no information regarding the reliability and validity of the BDIED (Bagnato, 1985; Carpenter, 1994). The BRIGANCE K & 1 Screen (Brigance, 1991) is a shorter version of the BDIED. Some other criterion-referenced assessments are Developmental Programming for Infants and Young

Children (Schafer & Moersch, 1981), The Hawaii Early Learning Profile (HELP) (Furuno et al., 1987), and The Early Learning Accomplishment Profile for Infants (Sanford, 1981).

The Infant Psychological Developmental Scale (Uzgiris & Hunt, 1975) is one of the most commonly used instruments based upon the Piagetian model. It assesses an infant's object-permanence ability, ability to understand means-ends and cause-effect relationships, ability to imitate vocalizations and gestures, and ability to manipulate objects in space.

The Metropolitan Readiness Test (*MRT,* Nurss & McGauvran, 1986) is currently in its fifth edition. This is a group-administered test measuring young children's fundamental competencies in reading, mathematics, and language-based activities. The administration of the test usually takes 80-90 minutes.

The MRT's Level I is designed for preschool children (4-6 year-olds). This level includes the following subtests: Auditory Memory, Beginning Consonants, Letter Reasoning, Visual Matching, School Language and Listening, and Quantitative Language. Level II is designed for kindergarten graduates and first-graders. The Level-II subtests are Beginning Consonants. Sound-Letter Correspondence, Visual Matching, Finding Patterns, School Language, Quantitative Concepts, and Quantitative Operations.

The Peabody Individual Achievement Test-Revised (*PIAT-R*, Markwardt, 1989) is a norm-referenced test (ages 5 to 18) of school achievement. It is administered individually and assesses performance in six content areas (general information, reading, recognition, reading comprehension, mathematics, spelling, and written expression). Level I of the PIAT-R is designed for kindergarteners and first graders and includes a number of prewriting skills (e.g., copying and writing letters). The PIAT-R was standardized on a representative sample of 1,563 students (K-12) and 175 kindergarteners. Most internal-consistency reliability indicators are over .90 with the exception of mathematics at the kindergarten level (.84). The criterion validity, as measured by correlations with a number of other tests (Sattler, 1992), is adequate.

Screening Devices

When large-scale assessments should be carried out, or when there is a need to determine which children may be developmentally at-risk and require further assessment, assessors may turn to certain special early child developmental screening devices. Such devices are somewhat predictive of scores from comprehensive assessments but require substantially less time to administer and score.

Due to their brevity, these instruments are neither as reliable nor as valid as comprehensive assessment tools. The psychometric properties of developmental screeners are defined by characteristics of the tests' sensitivity (the number of "misses," i.e., the degree of accuracy in detecting children with delays or disabilities) and specificity (the proportion of "false alarms," i.e., mislabeling children as delayed or disabled when in fact the children are developmentally normal). It has been recommended (Meisels, 1989) that both sensitivity and specificity levels of screening devices should be at least 80%. Although both indicators are important, it is usually assumed that sensitivity is a more critical characteristic of a screening device—follow-up assessments will correct false positives, whereas false negatives usually will not be referred for further assessment.

The Denver Developmental Screening Test—II. The Denver Developmental Screening Test—II (Frankenburg et al., 1990) is one of the most popular developmental screening tests (with an age range of 1 month to 6 years). The driving factor of the Denver is its brevity—it takes 15-20 minutes to

administer. The content of the test includes personal-social, motor, language, and adaptive domains. The scoring is based on parent reports, direct child assessment, and observation. The assessment result is expressed by a single score assigning the child to one of the four descriptive categories: Pass, Questionable, Abnormal, or Untestable.

The standardization of the Denver—II was done exclusively using young children from Colorado; therefore, the norms should be used cautiously.

The original version of the Denver (Frankenburg, Dodds, & Fandal, 1975) was found to be insufficiently sensitive to identify correctly most children with developmental delays or disabilities (Greer, Bauchner & Zuckerman, 1989). The sensitivity of the Denver—II has been greatly improved, but there is now evidence that it significantly over-identifies children as developmentally delayed or disabled (Galscoe & Byrne, 1993; Johnson, Ashford, Byrne, & Glascoe, 1992). The reported test-retest reliability is .90 and the reported inter-rater reliability is .98.

The Denver-II provides forms in Spanish, but no norms for children of diverse cultural backgrounds are available.

Other tests. The Early Screening Profile (ESP; Harrison, 1990) and the Developmental Indicators for the Assessment of Learning (DIAL-R, Mardell-Czudnowski & Goldberg, 1990) are examples of developmental screening instruments capitalizing primarily on direct assessment of the child. Both instruments are applicable to children 2- to 6-years-old, both take about 30 minutes to administer, both assess motor, language, and cognitive functioning (EPS also evaluates a child's personal-social and adaptive behaviors), both are normed on large, representative samples of children, and both exemplify some of the soundest validation available in the developmental-screening field. Test-retest reliability coefficients were .87 for the DIAL-R and .78-.89 for the ESP. Both instruments are available in English only.

In contrast to the ESP and DIAL-R, the Developmental Profile-II (DP-II; Alper, Boll, & Shearer, 1986) and the Developmental Observation Checklist System (DOCS; Hrescko, Miguel, Sherbenou, 1994) solely utilize data obtained from caregivers' reports. Both tools demonstrate adequate psychometric properties, evaluate children in a number of domains (DP-II: muscle and motor abilities, self-help, social, cognitive/intellectual, expressive and receptive communication; DOCS: language, motor, social, and cognitive development, child's adjustment behavior, and levels of family stress and support), and were standardized on large, nationally representative samples of children. The age-frame of the DOCS is 0 to 6 years and the age-frame of the DP-II is birth to 9 1/2 years.

The BRIGANCE Preschool Screen (Brigance, 1985) assesses children between 3 and 4 years of age. The administration takes 10-15 minutes. The reported internal-consistency reliability coefficient is .82 and the reported test-retest reliability indicator is .97. The device evaluates children's functioning in motor, language, body parts, colors, and personal domains. The standardization sample is small (408 children only). The test directions are available in English and Spanish.

The Miller Assessment for Preschoolers (MAP, Miller, 1988) is designed to screen children between 2.9 and 5.8 years of age. The administration of the MAP takes 25-30 minutes. The reported test-retest reliability ranges between .81 and .98. The MAP assesses children in three domains: motor, language, and cognition. The MAP is available in English only.

IV. Concluding Remarks

In closing this discussion, we would like to make three remarks.

First, there is no single instrument whose properties in all domains of comparison (qualitative and quantitative) stand out uniformly. For example, some instruments are psychometrically more sound, whereas others appear to be more developmentally appropriate, or include more engaging and more appealing materials. Some instruments are more theoretically grounded whereas others are primarily empirically driven. Some instruments are better in differentiating the upper tail of the cognitive ability distribution, whereas others better differentiate the lower tail.

Second, there is no unified rule regarding which test should be used when. Multiple issues should be considered when a decision is made in selecting a test. These issues should include (but are not be limited to) the purpose of testing, conditions of testing, the tester's expertise, the availability of materials, and the cost.

Finally, when a test is considered for use in a culture different from that where it was developed and standardized, an "implantation" in a different culture should be carried out very cautiously. Many (if not all) tests are culturally biased. They tend to favor the performance of children raised in the culture in which the test was created and suppress the performance of children from different cultures.

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile.* Burlington, VT: University of Vermont Department of Psychiatry.

Allen, S. R., & Thorndike, R. M. (1995). Stability of the WPPSI-R and WISC-III factor structure using corss-validation of covariance structure models. *Journal of Psychoeducational Assessment,* 13, 3-20.

Argulewicz, E. N., & Abel, R. R. (1984). Internal evidence of bias in the PPVT-R for Anglo-American and Mexican-American children. *Journal of School Psychology,* 22, 299-303.

Atkinson, L., Bevc, I., Dickens, S., & Blackwell, J. (1992). Concurrent validities of the Stanford-Binet (Fourth Edition), Leiter, and Vineland with developmentally delayed children. *Journal of School Psychology,* 30, 165-173.

Alfonso, V. C., & Flanagan, D. P. (1999). Assessment of cognitive functioning in preschoolers. In E. V. Nuttall, I. Ramero, & J. Kalesnik (Eds.), *Assessing and screening preschoolers* (pp. 186-217). Boston, MA: Allyn and Bacon.

Alpern, G. D., Boll, T. J., & Shearer, M. (1986). *Developmental Profile II.* Aspen, CO: Psychological Development Publication.

Bagnato, J. (1985). The BRIGANCE® Diagnostic Inventory of Early Development (BDIED). In J. V. Mitchell (Ed.), *The ninth mental measurement yearbook* (p.21). Lincoln, NE: University of Nebraska Press.

Bayley, N. (1993). *Bayley scales of infant development* (2nd ed.). San Antonio, TX: Psychological Corporation.

Boyd, R. D. (1989). What a difference a day makes: Age-related discontinuities and the Battelle Developmental Inventory. *Journal of Early Intervention,* 13, 114-119.

Boyd, R. D., Welge, P., Sexton, D., & Miller, J. H. (1989). Concurrent validity of the Battelle Developmental Inventory: Relationship with the Bayley Scales in young children with known or suspected disabilities. *Journal of Early Intervention,* 13, 14-23.

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychological Assessment,* 4, 313-326.

Brazelton, T. B. (1984). *Neonatal behavioral assessment scale* (2nd ed.). Clinics in Developmental Medicine (_ 88). Philadelphia: J. B. Lippincott.

Brigance, (1991). *Brigance K and 1 Screen for Kindergarten and First Grade.* North Billerica, MA; Curriculum Associates.

Brigance, A. H. (1991). *Brigance Diagnostic Inventory of Early Development: Revised.* North Billerica, MA: Curriculum Associates.

Brown, T L., & Morgan, S. B. (1991). Concurrent validity of the Stanford-Binet, 4th Edition: Agreement with the WISC—R in classifying learning disabled children. *Psychological Assessment,* 3, 247-253.

Bzoch, K., & League, R. (1991). *The receptive-expressive emergent language scale (REEL-2).* Austin, TX: Pro-Ed.

Carpenter, C. D. (1994). Review of the Revised BRIGANCE® Diagnostic Inventory of Early Development. In J. C. Connely & J. C. Impera (Eds.), *Supplement to the eleventh mental measurement yearbook* (pp. 352-353). Lincoln, NE: University of Nebraska Press.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, England: Cambridge University Press.

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. In. D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 122-130). New York: Guilford Press.

Carvajal, H., Hardy, K., Smith, K. L., & Weaver, K. A. (1988). Relationships between scores on Stanford-Binet IV and Wechsler Preschool and Primary Scale of Intelligence. *Psychology in the Schools,* 25, 129-131.

Cattell, P. (1960). *Cattell infant intelligence scale.* Cleveland, OH: Psychological Corporation.

Christensen, A.-L. (1984). *Luria's neuropsychological investigation* (2nd ed.). Copenhagen, Denmark, Munksgaard.

Colombo, J. (1993). *Infant cognition: Predicting later intellectual functioning.* Newbury Park, CA: Sage.

Conners, C. K. (1990). *Conners' Rating Scales manual.* North Tonawanda, NY: Multi-Health Systems.

Delaney, E., & Hopkins, T. (1987). *Examiner's handbook: An expanded guide for fourth edition users.* Chicago: Riverside.

Dietz, K. R., Lavigne, J. V., Arend, R., & Rosenbaum, D. (1997). Relation between intelligence and psychopathology among preschoolers. *Journal of Clinical Child Psychology, 26,* Mar 1997, 99-107.

Elliott, C. D. (1990a). *Differential Ability Scales: Administration and scoring manual.* San Antonio, TX: The Psychological Corporation.

Elliott, C. D. (1990b). *Differential Ability Scales: Introductory and technical handbook.* San Antonio, TX: The Psychological Corporation.

Flanagan, D. P., & Alfonso, V. C. (1995). A critical review of the technical characteristics of new and recently revised intelligence tests for preschool children. *Journal of Psychoeducational Assessment, 13,* 66-90.

Flanagan, D. P., Alfonso, V. C., Kaminer, T., & Rader, D. E. (1995). Incidence of basic concepts in the directions of new and recently revised American intelligence tests for preschoolers. *School Psychology International, 16,* 345-364.

Folio, M. R., & Fewell, R. R. (1983). *Peabody developmental motor scale and activity cards.* Allen, TX: Teaching Resources.

Frankenburg, W. K., Dodds, J., & Fandal. A. (1975). *Denver developmental screening test.* Denver, CO: LADOCA.

Frankenburg, W. K., Dodds, J., Archer, P., Bresnick, B., Maschka, P., Edelman, N., & Shapiro, H. (1990). *Denver II: Technical manual.* Denver, CO: Denver Developmental Materials.

Furuno, S., O'Reilly, K. A., Hosaka, C. M., Inatsuka, T. T., Allman, T. L., & Zeisloft, B. (1978). *Hawaii Early Learning Profile (HELP): Activity Guide.* Palo Alto, CA: VORT.

Gesell, A. (1940). The first five years of life: A guide to the study of the preschool child. New York: Harper.

Gibbs, E. D. (1990). Assessment of infant mental ability: Conventional tests and issues of prediction. In E. D. Gibbs & D. Teti (Eds.), *Interdisciplinary assessment of infants: A guide for early intervention professionals* (pp. 77-90). Baltimore, MD: Brookes.

Gilliam, W. S. (in press). Developmental assessment: Its role in the comprehensive psychiatric assessment of young children. In L. C. Mayes & W. C. Gilliam (Guest Eds.), "Comprehensive Psychiatric Assessment of Young Children" [Special Issue], *Child and Adolescent Psychiatric Clinics of North America.* Philadelphia: Saunders.

Gilliam, W. S., & Mayes, L. (in press). Developmental assessment of infants and toddlers. In C. H. Zeanah (Ed.), *Handbook of Infant Mental Health* (2nd ed.). New York: Guilford.

Glascoe, E. P., & Byrne, K. E. (1993). The accuracy of three developmental screening tests. *Journal of Early Intervention,* 17, 358-379.

Glutting, J., & Kaplan, D. (1990). Stanford-Binet Intelligence Scale: Fourth Edition: Making the case for reasonable interpretations. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 277-296). New York: Guilford Press.

Greer, S. Bauchner, H., & Zuckerman, B. (1989). The Denver Developmental Screening Test: How good is its predictive validity? *Developmental Medicine and Child Neurology, 31,* 774-781.

Griffiths, R. (1954). *The abilities of babies.* London: University of London Press.

Griffiths, R. (1979). *The abilities of young children.* London: Child Development Research Center.

Gyurke, J. S. (1991). The assessment of preschool children with the Wechsler Preschool and Primary

Scale of Intelligence-Revised. In B. A. Bracken (Ed.), *The psychoeducational assessment of preschool children* (2nd ed., pp. 86-106). Boston: Allyn & Bacon.

Harrington, R. G., Kimbrell, J., & Dai, X. (1992). The relationship between the Woodcock-Johnson Psycho-Educational Battery—Revised (Early Development) and the Wechsler Preschool and Primary Scale of Intelligence—Revised. *Psychology in the Schools, 29,* 116-125.

Harrison, P. L. (1990). *Early Screening Profiles (ESP): Manual.* Circle Pines, MN: American Guidance.

Hessler, G. (1993). *Use and interpretation of the Woodcock-Johnson Psycho-Educational Battery-Revised.* Chicago: Riverside.

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock (Eds.), *WJ-R technical manual* (pp. 197-223). Chicago: Riverside.

Horn, J. L. (1994). Theory of fluid and crystallized intelligence. In R. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 443-451). New York: Macmillan.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53-91). New York: Guilford Press.

Horowitz, F. D. & Linn, L. P. (1982). The Neonatal Behavioral Assessment Scale. In M. Wolraich & D. K. Routh (Eds.), *Advances in Developmental Pediatrics, 3* (pp. 223-256). Greenwich, CT: JAI.

Hrescko, W. P., Miguel, S. A., Sherbenou, R. J., et al. (1994). *Developmental Observational Checklist System.* Austin, TX: Pro-Ed.

Johnson, K. L., Ashford, L. G., Byrne, K. E., & Glascoe, F. P. (1992). Does Denver II produce meaningful results? [Letter to the Editor]. *Pediatrics, 90,* 477-478.

Kaplan, S. L., & Alfonso, V. C. (1997). Confirmatory factor analysis of the Stanford-Binet Intelligence Scale: Fourth Edition with preschoolers with developmental delays. *Journal of Psychoeducational Assessment, 15,* 226-236.

Karr, S. K., Carvajal, H., & Palmer, B. L. (1992). Comparison of Kaufman's short form of the McCarthy Scales of Children's Abilities and the Stanford-Binet Intelligence Scales—Fourth Edition. *Perceptual & Motor Skills, 74,* 1120-1122.

Keith, T. Z. (1990). Confirmatory and hierarchical confirmatory analysis of the Differential Ability Scales. *Journal of Psychoeducational Assessment, 8,* 391-405.

Keith, T. Z., Coll, V. A., Novak, C. G., White, L. J., & Pottebaum, S. M. (1988). Confirmatory factor analysis of the Stanford-Binet Fourth Edition: Testing the theory-test match. *Journal of Visual Impairment and Blindness, 90,* 423-436.

Kline, R. B. (1989). Is the Fourth Edition Stanford Binet a four-factor test? Confirmatory factor analyses of alternative models for ages 2 through 23. *Journal of Psychoeducational Assessment, 7,* 4-13.

Kline, R. B., Snyder, J., Guilmette, S., & Castellanos, M. (1993). External validity of the profile variability index for the K-ABC, Stanford-Binet, and WISC—R: Another cul-de-sac. *Journal of Learning Disabilities, 26,* 557-567.

Knoff, , H. M., Stollar, S. A., Johnson, J. J., & Chenneville, T. A. (1999). Assessment of social-emotional functioning and adaptive behavior. In E. V. Nuttall, I. Ramero, & J. Kalesnik (Eds.), *Assessing and screening preschoolers* (pp. 126-160). Boston, MA: Allyn and Bacon.

Korkman, M., Kirk, U., & Kemp, S. (1998). *NEPSY: A developmental neuropsychological assessment.* San Antonio, TX: The Psychological Corporation.

Lamp, R. E., & Krohn, E. J. (1990). Stability of the Stanford-Binet Fourth Edition and K-ABC for young Black and White children from low income families. *Journal of Psychoeducational Assessment, 8,* 139-149.

Lipsitt, L. P. (1992). Discussion: The Bayley scales of infant development: Issues of prediction and outcome revisited. *Advances in Infancy Research, 7,* 239-245.

Luria, A. R. (1973). *The working brain: An introduction to neuropsychology.* London: Penguin.

Mardell-Czudnowski, C. D., Goldberg, D. (1990). *DIAL-R (Developmental Indicators for the Assessment of Learning—Revised).* Edison, NJ: Childcraft Education.

Markwardt, F. C. (1989). *Peabody Individual Achievement Test—Revised.* Circle Pines, MN: American Guidance Service.

Matula, K., Gyurke, J. S., & Aylward, G. P. (1997). Response to commentary. Bayley Scales-II, *Developmental and Behavioral Pediatrics,* 18, 112-113.

McCarth, D. A. (1978). *Manual for the McCarthy Scales of Children's Abilities.* San Antonio: The Psychological Corporation.

McCrowell, K. L., & Nagle, R. J. (1994). Comparability of the WPPSI—R and the S-B:IV among preschool children. *Journal of Psychoeducational Assessment, 12,* 126-134.

McGrew, K. S., and Flanagan, D. P. (1996). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment.* Boston: Allyn & Bacon.

McLinden, S. E. (1989). An evaluation of the Battelle Developmental Inventory for determining special educational eligibility. *Journal of Psychoeducational Assessment, 7,* 66-73.

Meisels, S. J., (1989). Can developmental screening tests identify children who are developmentally at risk? *Pediatrics, 83,* 578-585.

Meisels, S. J., & Waskik, B. A. (1990). Who should be served? Identifying children in need of early intervention. In S. J. Meisels & J. P. Shonkoff (Eds.), *Handbook of early childhood intervention* (pp. 605-632). New York: Cambridge.

Miller, L. J. (1988). *Miller Assessment for Preschoolers.* San Antonio, TC: The Psychological Corporation.

Molfese, V., Yaple, K., Helwig, S., Harris, L., & Connell, S. (1992). Stanford-Binet Intelligence Scale (4ht ed.): Factor structure and verbal subscale scores for three-year-olds. *Journal of Psychoeducational Assessment, 10,* 47-58.

Mullen, E. M. (1995). *Mullen scales of early learning: AGS edition.* Circle Pines, MN: American Guidance Service.

Newborg, J., Stock, J., Wnek, L., Guidubaldi, J., & Svinicki, J. S. (1984). *Battelle Developmental Inventory (BDI).* Allen, TX: DLM/Teaching Resources.

Newland, T. E. (1971). Psychological assessment of exceptional children and youth. In W. Cruickshank (Ed.), *Psychology of exceptional children and youth* (pp. 115-172). Englewood Cliffs, NJ: Prentice-Hall.

Nurss, J., & McGauvran, M. (1986). *Metropolitan Readiness Assessment Program.* San Antonio, TX: The Psychological Corporation.

Ownby, R. L., & Carmin, C. N. (1988). Confirmatory factor analysis of the Stanford-Binet Intelligence Scale-Fourth Edition. *Journal of Psychoeducational Assessment, 6,* 331-340.

Phelps, L., Bell, M. C., & Scott, M. J. (1988). Correlations between the Stanford-Binet: Fourth Edition and the WISC—R with a learning disabled population. *Psychology in the Schools, 25,* 380-382.

Reynolds, C. R., Wilson, V. L., & Chatman, S. P. (1984). Item bias on the 1981 revision of the Peabody Picture Vocabulary Test using a new method of detecting bias. *Journal of Psychoeducational Assessment, 2,* 219-224.

Ross, G., & Lawson, K. (1997a). The Graham-Rosenblith behavioral examination for newborns: Prognostic values and procedural issues, In J. Osofsky (Ed.), *Handbook of infant development.* New York: Wiley.

Rossetti, L. M. (1990). *Infant-toddler assessment: An interdisciplinary approach.* Boston: College Hill.

Rothlisberg, B A., & McIntosh, D. E. (1991). Performance of a referred sample on the Stanford-Binet IV and the K-ABC. *Journal of School Psychology, 29,* 367-370.

Salvia, J. & Ysseldyke, J. E. (1991). *Assessment* (5th ed.). Boston: Houghton Mifflin.

Sameroff, A. J. (1978). Organization and stability of newborn behavior: A commentary on the Brazelton Neonatal Behavior Assessment Scale. *Monographs of the Society for Research in Child Development, 43.*

Sanford, A. (1981). *Learning Accomplishment Profile for Infants (Early LAP).* Winston-Salem, NC: Kaplan School Supply.

Sattler, J. M. (1992). *Assessment of children* (3rd ed.-Revised). San Diego, CA: Author.

Sparrow, S. S., Balla, B. D., & Cicchetti, D. V. (1998). *Vineland Social-Emotional Early Childhoos Scales.* Circle Pines, MN: American Guidance Service.

Stone, B. J., Gridley, B., & Gyurke, J. S. (1991). Confirmatory factor analysis of the WPPSI-R at the extreme end of the age range. *Journal of Psychoeducational Assessment, 9,* 263-270.

Thomas, H. (1970). Psychological assessment instruments for use with human infants. *Merrill-Palmer Quarterly, 16,* 179-223.

Thorndike, R., Hagen, E., & Sattler, J. R. (1986a). *Guide for administering and scoring the Stanford-Binet Intelligence Scale* (4th ed, 2nd pr.). Chicago, IL: Riverside.

Thorndike, R., Hagen, E., & Sattler, J. R. (1986b). *Technical Manual. Stanford-Binet Intelligence Scale* (4th ed). Chicago, IL: Riverside.

Uzgis, I., & Hunt, J. (1975). *Assessment in infancy: Ordinal scales of psychological development.* Urbana: University of Illinois.

Wechsler, D. (1949). *Manual for the Wechsler Intelligence Scale for Children.* New York: The Psychological Corporation.

Wechsler, D. (1967). *Wechsler Preschool and Primary Scale of Intelligence.* New York: Psychological Corporation.

Wechsler, D. (1989). *Manual for the Wechsler Preschool and Primary Scale of Intelligence-Revised.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children-III.* San Antonio: The Psychological Corporation.

Whatley, J. (1987). Bayley Scales of Infant Development. In D. Keyse & R. Sweetland (Eds.), *Test critique* (Vol 6, pp. 38-47). Kansas City, MO: Westport.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery-Revised.* Allen, TX: DLM.

Woodcock, R. W., & Mather, N. (1989). *The WJ-R tests of cognitive ability standard and supplemental batteries: Examiner's manual.* Chicago, IL: Riverside.

Woodcock, R. W., & Muñoz-Sandoval, A. F. (1996). *Batería Woodcock-Muñoz: pruebas de habilidad cognitive—revisada.* Chicago, IL: Riverside.

APPENDIX

Additional References

The Bayley Scales of Infant Development

■ First and Second Editions: A Comparison

DeWitt, M. B., Schreck, K. A., Mulick, J. A. (1998). Use of Bayley Scales in individuals with profound mental retardation: Comparison of the first and second editions. *Journal of Developmental & Physical Disabilities, 10,* 307-313.

Goldstein, D. J, Fogle, E. E, Wieber, J. L, O'Shea, T. M. (1995). Comparison of the Bayley Scales of Infant Development-Second Edition and the Bayley Scales of Infant Development with premature infants. *Journal of Psychoeducational Assessment, 13,* 391-396.

Nellis, L., Gridley, B. E. (1994). Review of the Bayley Scales of Infant Development—Second Edition. *Journal of School Psychology, 32,* 201-209.

■ Bayley Scales-II: Description, Administration, Scoring, Interpretation

Kaplan-Estrin, M., Jacobson, S. W., & Jacobson, J. L. (1994). Alternative approaches to clustering and scoring the Bayley Infant Behavior Record. *Infant Behavior & Development, 17,* 149-157.

Robinson, B. F., Mervis, C. B. (1996). Extrapolated raw scores for the second edition of the Bayley Scales of Infant Development. *American Journal on Mental Retardation, 100,* 666-670.

Sibian, A. H. (1994) *Sequences of perceptual, social, cognitive and language processes as measured by the Bayley Mental Scale of Infant Development.* Dissertation Abstracts International Section A: Humanities & Social Sciences, Vol 54(9-A), 1994, 3384.

Siegel, L. S., Cooper, D. C., Fitzhardinge, P. M., & Ash, A. J. (1995). The use of the Mental Development Index of the Bayley Scale to diagnose language delay in 2-year-old high risk infants. *Infant Behavior & Development, 18,* 483-486.

Vandermeulen, B. F., Smrkovsky, M., Lecoultre-Martin, P., & Wijnberg-Williams, B. J. (1994). A nonverbal version of the Bayley Scales of Infant Development. *Psychologica Belgica, 34,* 141-152.

■ Psychometric Properties

Aylward, G. P., Verhulst, S. J., Bell, S., Gyurke, J. S, et al. (1995). Cognitive and motor score differences in biologically at-risk infants. Infant Behavior & Development, 18, 43-52.

Costarides, A. H., Shulman, B. B. (1998). Norm-referenced language measures: Implications for assessment of infants and toddlers. *Topics in Language Disorders, 18,* 26-33.

Fagen, J. W., Singer, J. M., Ohr, P. S., & Fleckenstein, L. K. (1987). Infant temperament and performance on the Bayley Scales of Infant Development at 4, 8, and 12 months of age. *Infant Behavior & Development, 10,* 505-512.

Gerken, K. C., Eliason, M. J., Arthur, C. R. (1994). The assessment of at-risk infants and toddlers with the Bayley Mental Scale and the Battelle Developmental Inventory: Beyond the data. *Psychology in the Schools, 31,* 181-187.

Molfese, V. J., & Acheson, S. (1997). Infant and preschool mental and verbal abilities: How are infant scores related to preschool scores? *International Journal of Behavioral Development, 20,* 595-607.

Molfese, V. J, DiLalla, L. F., & Lovelace, L. (1996). Perinatal, home environment, and infant measures as successful predictors of preschool cognitive and verbal abilities. *International Journal of Behavioral Development, 19,* 101-119.

Raggio, D. J., Massingale, T. W., & Bass, J. D. (1994). Comparison of Vineland Adaptive Behavior Scales-Survey Form age equivalent and standard score with the Bayley Mental Development Index. *Perceptual & Motor Skills, 79,* 203-206.

Raggio, D. J., Massingale, & Twila W. (1993). Comparison of the Vineland Social Maturity Scale, The Vineland Adaptive Behavior Scales—Survey Form, and the Bayley Scales of Infant Development with infants evaluated for developmental delay. *Perceptual & Motor Skills, 77,* 931-937.

Rossman, M. J., Hyman, S. L., Rorabaugh, M. L., Berlin, L. E., et al. (1994). The CAT/CLAMS assessment for early intervention services. *Clinical Pediatrics, 33,* 404-409.

Thompson, B., &Wasserman, J. D., & Matula, K. (1996). The factor structure of the Behavior Rating Scale of the Bayley Scales of Infant Development-II. *Educational & Psychological Measurement, 56,* 460-474.

■ Special Populations

Hancock, F. A. (1993) *The relationship of familiarity with examiner, birth status, and sociability to measured developmental level of preterm and full-term infants.* Dissertation Abstracts International. Vol 54(5-A), 1993, 1665.

Pryor, J. (1996) Physical and behavioural correlates of 12-month development in small-for-gestational age and appropriately grown infants. *Journal of Reproductive & Infant Psychology, 14,* 233-242.

Ross, G., Lawson, K. (1997b). Using the Bayley-II: Unresolved issues in assessing the development of prematurely born children. *Journal of Developmental & Behavioral Pediatrics, 18,* 109-111.

Siegel, L. S., Cooper, D. C., Fitzhardinge, P. M., & Ash, A. J. (1995). The use of the Mental Development Index of the Bayley Scale to diagnose language delay in 2-year-old high risk infants. *Infant Behavior & Development, 18,* 483-486.

Canter, A. (1990). A new Binet, and old premise: A mismatch between technology and evolving practice. *Journal of Psychoeducational Assessment, 8,* 443-450.

Choi, H., & Proctor, T. B. (1994). Error-prone subtests and error types in the administration of the Stanford-Binet Intelligence Scale: Fourth edition. *Journal of Psychoeducational Assessment, 12,* 165-171.

Glutting, J. J.(1989). Introduction to the structure and application of the Stanford-Binet Intelligence Scale—Fourth Edition. *Journal of School Psychology, 27,* 69-80.

Kramer, J. J; Henning-Stout, Mary; Ullman, Daniel P; Schellenberg, Richard P.(1987) The viability of scatter analysis on the WISC—R and the SBIS: Examining a vestige. *Journal of Psychoeducational Assessment. Vol 5*(1), 37-47.

Molfese, V. J., Helwig, S., Holcomb, L. (1993) Standardized assessments of verbal intelligence in 3-year-old children: A comparison of biomedical and psychoeducational data in a longitudinal sample. *Journal of Psychoeducational Assessment, 11,* 56-66.

Naglieri, J. A. (1988). Interpreting area score variation on the fourth edition of the Stanford-Binet Scale of Intelligence. *Journal of Clinical Child Psychology, 17,* 225-228.

Rosenthal, B. L., & Kamphaus, R. W. (1988). Interpretive tables for test scatter on the Stanford-Binet Intelligence Scale: Fourth Edition. *Journal of Psychoeducational Assessment, 6,* 359-370.

Sabatino, D. A. (1993). Ascertaining intellectual functioning with Binet-type instruments. In V. H. Booney (Ed). *Best practices in assessment for school and clinical settings,* (pp. 147-175). Brandon, VT, USA: Clinical Psychology Publishing Co, Inc.

Spruill, J. (1988). Two types of tables for use with the Stanford Binet Intelligence Scale: Fourth Edition. *Journal of Psychoeducational Assessment, 6,* 78-86.

Thorndike, R. L. (1986). Bayesian concepts and test making. *Journal of Counseling & Development, 65,* 110-111.

Vernon, P. E. (1987). The demise of the Stanford-Binet Scale. *Canadian Psychology, 28,* 251-258.

Wersh, J., & Thomas, M. R. (1990). The Stanford-Binet Intelligence Scale—Fourth Edition: Observations, comments and concerns. *Canadian Psychology, 31,* 190-193.

■ Previous Editions, Concurrent Versions, and Alternative Forms

Atkinson, L. (1991). Short forms of the Stanford-Binet Intelligence Scale, Fourth Edition, for children with low intelligence. *Journal of School Psychology, 29,* 177-181.

DeLamatre, J. E., & Hollinger, C. L. (1990). Utility of the Stanford-Binet IV abbreviated form for placing exceptional children. *Psychological Reports, 67,* 973-974.

Glaub, V. E., & Kamphaus, R. W. (1991). Construction of a nonverbal adaptation of the Stanford-Binet Fourth Edition. *Educational & Psychological Measurement, 51,* 231-241.

Husband, T. H., & Hayden, D. C. (1996). Effects of the addition of color to assessment instruments. *Journal of Psychoeducational Assessment, 14,* 147-151.

Kitano, M. K., de Leon, J. (1988) Use of the Stanford Binet Fourth Edition in identifying young gifted children. *Roeper Review, 10,* 156-159.

Kluever, R. C., Green, K. E. (1990). Identification of gifted children: A comparison of the scores on Stanford-Binet 4th Edition and Form LM. *Roeper Review, 13,* 16-20.

Kyle, J. M., & Robertson, C. M. T. (1994). Evaluation of three abbreviated forms of the Stanford-Binet Intelligence Scale: Fourth edition. *Canadian Journal of School Psychology, 10,* 147-154.

Lawson, T. T., & Evans, L. D. (1996). Stanford-Binet: Fourth edition short forms with underachieving and learning disabled students. *Psychological Reports, 79,* 47-50.

McCall, V. W., Yates, B., Hendricks, S., Turner, K. et al. (1989) Comparison between the Stanford-Binet: L-M and the Stanford-Binet: Fourth Edition with a group of gifted children. *Contemporary Educational Psychology, 14,* 93-96.

Prewett, P. N. (1992) Short forms of the Stanford-Binet Intelligence Scale: Fourth Edition. *Journal of Psychoeducational Assessment. Vol 10*(3), 257-264.

Riccio, C. A., Platt, L. O., K., Randy, W., Greer, M. K., et al. (1994) Principal components analysis of the General Purpose Abbreviated Battery of the Stanford-Binet, Fourth Edition, for young children. *Assessment, 1,* 173-178.

Robinson, N. M. (1992). Which Stanford-Binet for the brightest? Stanford-Binet IV, of course? Time marches on. *Roeper Review, 15,* 32-34.

Robinson, N. M., Dale, P. S., & Landesman, S. (1990). Validity of Stanford-Binet IV with linguistically precocious toddlers. *Intelligence, 14,* 173-186.

■ Racial Differences

Montie, J. E., Fagan, J. F. (1993). Racial differences in IQ: Item analysis of the Stanford-Binet at 3 years. *Intelligence, 19,* 315-332.

Peoples, C. E., Fagan, J. F., Drotar, D. (1995). The influence of race on 3-year-old children's performance on the Stanford-Binet: Fourth Edition. *Intelligence, 21,* 69-82.

■ SES Differences

Krohn, E. J., & Lamp, R. E. (1989). Concurrent validity of the Stanford-Binet Fourth Edition and K-ABC for Head Start children. *Journal of School Psychology, 27,* 59-67.

Lamp, R. E., & Krohn, E. J. (1990). Stability of the Stanford-Binet Fourth Edition and K-ABC for young Black and White children from low income families. *Journal of Psychoeducational Assessment, 8,* 139-149.

Prewett, P. N., & Matavich, M. A. (1994). A comparison of referred students' performance on the WISC-III and the Stanford-Binet Intelligence Scale: Fourth Edition. *Journal of Psychoeducational Assessment, 12,* 42-48.

■ Special Populations

Atkinson, L. (1991). Short forms of the Stanford-Binet Intelligence Scale, Fourth Edition, for children with low intelligence. *Journal of School Psychology. Vol 29*(2), 177-181.

Bower, A., & Hayes, A. (1995). Relations of scores on the Stanford Binet fourth edition and form L-M: Concurrent validation study with children who have mental retardation. *American Journal on Mental Retardation, 99,* 555-558.

Brown, T. L., & Morgan, S. B. (1991). Concurrent validity of the Stanford-Binet, 4th Edition: Agreement with the WISC—R in classifying learning disabled children. *Psychological Assessment, 3,* 247-253.

DeLamatre, J. E., & Hollinger, C L. (1990). Utility of the Stanford-Binet IV abbreviated form for placing exceptional children. *Psychological Reports, 67,* 973-974.

Glaub, V. E., & Kamphaus, R. W. (1991). Construction of a nonverbal adaptation of the Stanford-Binet Fourth Edition. *Educational & Psychological Measurement, 51,* 231-241.

Harris, S. L., Handleman, J. S., & Burton, J. L. (1990). The Stanford Binet profiles of young children with autism. *Special Services in the Schools, 6,* 135-143.

Knight, B. C, Baker, E. H., & Minder, C. C. (1990). Concurrent validity of the Stanford-Binet: Fourth Edition and Kaufman Assessment Battery for Children with learning-disabled students. *Psychology in the Schools, 27,* 116-125.

Lavin, C. (1995) Clinical applications of the Stanford-Binet Intelligence Scale: Fourth Edition to reading instruction of children with learning disabilities. *Psychology in the Schools, 32,* 255-263.

Phelps, L., Bell, M. C., & Scott, M. J. (1988). Correlations between the Stanford-Binet: Fourth Edition and the WISC—R with a learning disabled population. *Psychology in the Schools, 25,* 380-382.

Robinson, N. M., Dale, P. S., & Landesman, S. (1990). Validity of Stanford-Binet IV with linguistically precocious toddlers. *Intelligence, 14,* 173-186.

Silverman, L K., & Kearney, K. (1992). Which Stanford-Binet for the brightest? The case for the Stanford-Binet L-M as a supplemental test. *Roeper Review, 15,* 34-37.

Spruill, J. (1996). Composite SAS of the Stanford-Binet Intelligence Scale, Fourth Edition: Is it determined by only one are SAS? *Psychological Assessment, 8,* 328-330.

Vig, S., & Jedrysek, E. (1996). Stanford-Binet Fourth Edition: Useful for young children with language impairment? *Psychology in the School, 33,* 124-131.

Wilson, W. M. (1992). The Stanford-Binet: Fourth Edition and Form L-M in assessment of young children with mental retardation. *Mental Retardation, 30,* 81-84.

■ Psychometric Properties

Atkinson, L. (1989). Three standard errors of measurement and the Stanford-Binet Intelligence Scale, Fourth Edition. *Psychological Assessment, 1,* 242-244.

Boyle, G. J. (1989). Confirmation of the structural dimensionality of the Stanford-Binet Intelligence Scale (fourth edition). *Personality & Individual Differences, 10,* 709-715.

Boyle, G. J. (1990) Stanford-Binet IV Intelligence Scale: Is its structure supported by LISREL congeneric factor analyses? *Personality & Individual Differences, 11,* 1175-1181.

Carvajal, H., Hardy, K., Harmon, K., Sellers, T. A., et al. (1987). Relationships among scores on the Stanford-Binet IV, Peabody Picture Vocabulary Test—Revised, and Columbia Mental Maturity Scale. *Bulletin of the Psychonomic Society, 25,* 275-276.

Carvajal, H., Karr, S. K., Hardy, K. M., & Palmer, B. L. (1988). Relationships between scores on Stanford-Binet IV and scores on McCarthy Scales of Children's Abilities. *Bulletin of the Psychonomic Society, 26,* 349.

Carvajal, H. H., Parks, J. P., Bays, K. J., Logan, R. A., et al. (1991). Relationships between scores on Wechsler Preschool and Primary Scale of Intelligence—Revised and Stanford-Binet IV. *Psychological Reports, 69,* 23-26.

Gerken, K. C., & Hodapp, A. F. (1992). Assessment of preschoolers at-risk with the WPPSI—R and the Stanford-Binet L-M. *Psychological Reports, 71,* 659-664.

Goldstein, D. J., & Sheaffer, C. I. (1988). Ratio developmental quotients from the Bayley are comparable to later IQs from the Stanford-Binet. *American Journal on Mental Retardation, 92,* 379-380.

Gridley, B. E., & McIntosh, D. E. (1991). Confirmatory factor analysis of the Stanford-Binet: Fourth Edition for a normal sample. *Journal of School Psychology, 29,* 237-248.

Hendershott, J. L., Searight, H. R., Hatfield, J. L., & Rogers, B J. (1990). Correlations between the Stanford-Binet, Fourth Edition and the Kaufman Assessment Battery for Children for a preschool sample. *Perceptual & Motor Skills, 71,* 819-825.

Hodapp, A. F. (1993). Correlation between Stanford-Binet IV and PPVT—R scores for young children. *Psychological Reports, 73,* 1152-1154.

Howell, K K; Bracken, B A. (1992). Clinical utility of the Bracken Basic Concept Scale as a preschool intellectual screener: Comparison with the Stanford-Binet for African-American children. *Journal of Clinical Child Psychology, 21,* 255-261.

Johnson, D. L., Howie, V. M., Owen, M., Baldwin, C. D. et al. (1993) Assessment of three-year-olds with the Stanford-Binet Fourth Edition. *Psychological Reports, 73,* 51-57.

Keith, T. Z., Cool, V. A., Novak, C. G., White, L. J., et al. (1988). Confirmatory factor analysis of the Stanford-Binet Fourth Edition: Testing the theory-test match. *Journal of School Psychology, 26,* 253-274.

Kline, R. B. (1989). Is the Fourth Edition Stanford-Binet a four-factor test? Confirmatory factor analyses of alternative models for ages 2 through 23. *Journal of Psychoeducational Assessment, 7,* 4-13.

Kline, R B., Snyder, J., Guilmette, S., & Castellanos, M. (1992). Relative usefulness of elevation, variability, and shape information from WISC—R, K-ABC, and Fourth Edition Stanford-Binet profiles in predicting achievement. *Psychological Assessment, 4,* 426-432.

Knight, B. C., Baker, E. H., & Minder, C. C. (1990). Concurrent validity of the Stanford-Binet: Fourth Edition and Kaufman Assessment Battery for Children with learning-disabled students. *Psychology in the Schools. Vol 27*(2), 116-125.

Krohn, E. J., & Lamp, R. E. (1989). Concurrent validity of the Stanford-Binet Fourth Edition and K-ABC for Head Start children. *Journal of School Psychology, 27,* 59-67.

Kunen, S., Overstreet, S., & Salles, C. (1996). Concurrent validity study of the Slosson Intelligence Test-Revised in mental retardation testing. *Mental Retardation, 34,* 380-386.

Laurent, J., Swerdlik, M., & Ryburn, M. (1992). Review of validity research on the Stanford-Binet Intelligence Scale: Fourth Edition. *Psychological Assessment, 4,* 102-112.

Mason, E M. (1992). Percent of agreement among raters and rater reliability of the Copying subtest of the Stanford-Binet Intelligence Scale: Fourth Edition. *Perceptual & Motor Skills, 74,* 347-353.

McCallum, R. S. (1990). Determining the factor structure of the Stanford-Binet—Fourth Edition: The right choice. *Journal of Psychoeducational Assessment, 8,* 436-442.

Molfese, V. J., & Acheson, S. (1997). Infant and preschool mental and verbal abilities: How are infant scores related to preschool scores? *International Journal of Behavioral Development, 20,* 595-607.

Naglieri, J. A. (1988). Interpreting the subtest profile on the fourth edition of the Stanford-Binet Scale of Intelligence. *Journal of Clinical Child Psychology, 17,* 62-65.

Ownby, R. L., & Carmin, C. N. (1988). Confirmatory factor analyses of the Stanford-Binet Intelligence Scale, Fourth Edition. *Journal of Psychoeducational Assessment, 6,* 331-340.

Prewett, P. N., & Farhney, M. R. (1994). The concurrent validity of the Matrix Analogies Test-Short Form with the Stanford-Binet: Fourth Edition and KTEA-BF (academic achievement). P*sychology in the Schools, 31,* 20-25.

Reynolds, C. R., Kamphaus, R. W., & Rosenthal, B. L. (1988). Factor analysis of the Stanford-Binet Fourth Edition for ages 2 years through 23 years. *Measurement & Evaluation in Counseling & Development, 21,* 52-63.

Riccio, C. A., Platt, L. O. K., Randy, W., Greer, M. K., et al. (1994). Principal components analysis of the General Purpose Abbreviated Battery of the Stanford-Binet, Fourth Edition, for young children. *Assessment, 1,* 173-178.

Robinson, N. M., Dale, P. S., & Landesman, S. (1990). Validity of Stanford-Binet IV with linguistically precocious toddlers. *Intelligence, 14,* 173-186.

Thorndike, R. M. (1990). Would the real factors of the Stanford-Binet Fourth Edition please come forward? *Journal of Psychoeducational Assessment, 8,* 412-435.

Schuerger, J. M., & Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology, 45,* 294-302.

Shanahan, R., & Bradley-Johnson, S.(1992). Concurrent validity of the Cognitive Abilities Scale and Binet IV for nonvocal 2- and 3-year-olds. *Journal of School Psychology, 30,* 395-399.

The Kaufman Assessment Battery for Children (K-ABC)

■ Test Description, Administration, and Interpretation

Applegate, B., & Kaufman, A. S.(1989). Short form of K-superABC sequential and simultaneous processing for research and screening. *Journal of Clinical Child Psychology, 18,* 305-313.

Bracken, B. A., & Fagan, T. K.(1988). Abilities assessed by the K-ABC Mental Processing subtests: The perceptions of practitioners with varying degrees of experience. *Psychology in the Schools, 25,* 22-34.

Bracken, B. A., & Howell, K. K. (1989). K-ABC subtest specificity recalculated. *Journal of School Psychology, 27,* 335-345.

Conoley, J. C. (1990). Review of the K-ABC: Reflecting the unobservable. *Journal of Psychoeducational Assessment, 8,* 369-375.

Glutting, J. J., McGrath, E. A., Kamphaus, R. W., & McDermott, P A.(1992). Taxonomy and validity of subtest profiles on the Kaufman Assessment Battery for Children. *Journal of Special Education, 26,* 85-115.

Gordon, M., Thomason, D., & Cooper, S. (1990). To what extent does attention affect K-ABC scores? *Psychology in the Schools, 27,* 144-147.

Hunnicutt, L. C., Slate, J. R., Gamble, C., & Wheeler, M. S. (1990). Examiner errors on the Kaufman Assessment Battery for Children: A preliminary investigation. *Journal of School Psychology, 28,* 271-278.

Kalmar, K., Massoth, N. A., & Westerveld, M.(1988). K-ABC mental processing score variations among normally functioning school-age children who display torque circle-drawing directionality. *International Journal of Clinical Neuropsychology, 10,* 97-102.

Kamphaus, R. W. (1990). K-ABC theory in historical and current contexts. *Journal of Psychoeducational Assessment, 8,* 356-368.

Kamphaus, R. W., & Reynolds, C. R. (1987). *Clinical and research applications of the K-ABC.* Circle Pines, MN, USA: American Guidance Service.

Kaufman, A. S., & Applegate, B. (1988). Short forms of the K-ABC Mental Processing and Achievement scales at ages 4 to 12_ years for clinical and screening purposes. *Journal of Clinical Child Psychology. Vol 17*(4), 359-369.

Kaufman, A. S., O'Neal, M. R., Avant, A. H., & Long, S. W. (1987). Introduction to the Kaufman Assessment Battery for Children (K-ABC) for pediatric neuroclinicians. *Journal of Child Neurology, 2,* 3-16.

Kline, R. B., Snyder, J., & Castellanos, M. (1996). Lessons from the Kaufman Assessment Battery for Children (K-ABC): Toward a new cognitive assessment model. *Psychological Assessment, 8,* 7-17.

Kline, R. B., Snyder, J., Guilmette, S., & Castellanos, M. (1992). Relative usefulness of elevation, variability, and shape information from WISC—R, K-ABC, and Fourth Edition Stanford-Binet profiles in predicting achievement. *Psychological Assessment, 4,* 426-432.

Kline, R. B., Snyder, J., Guilmette, S., & Castellanos, M. (1994). Evaluation of the construct validity of the Kamphaus-Reynolds supplementary scoring system for the K-ABC. *Assessment, 1,* 219-226.

Levenson, R. L. (1985). Kaufman Assessment Battery for Children: Alternate solutions to triangles, Item 17. *Perceptual & Motor Skills, 61,* 73-74.

Reynolds, C. R., & Kamphaus, R W. (1986). The Kaufman Assessment Battery for Children: Development, structure, and application in neuropsychology. In D. Wedding & A. Horton (Eds), *The neu-*

ropsychology handbook: Behavioral and clinical perspectives. (pp. 194-216). New York, NY: Springer Publishing Co, Inc.

Reynolds, C. R., Kamphaus, R. W., & Rosenthal, B. L. (1989). Applications of the Kaufman Assessment Battery for Children (K-ABC) in neuropsychological assessment. In C. R. Reynolds. & E. Fletcher-Janzen (Eds), *Handbook of clinical child neuropsychology. Critical issues in neuropsychology.* (pp. 205-226). New York, NY: Plenum Press.

Reynolds, C. R., Kamphaus, R. W. (1997). The Kaufman Assessment Battery for Children: Development, structure, and applications in neuropsychology. In A. Horton, & D. Wedding (Eds), *The neuropsychology handbook, Vol. 1: Foundations and assessment (2nd ed.).* (pp. 290-330). New York, NY: Springer Publishing Co, Inc.

■ Special Populations

Bain, S. K. (1993). Sequential and simultaneous processing in children with learning disabilities: An attempted replication. *Journal of Special Education, 27,* 235-246.

Donders, J. (1992). Validity of the Kaufman Assessment Battery for Children when employed with children with traumatic brain injury. *Journal of Clinical Psychology. Vol 48*(2), 225-230.

Gibbins, S. (1989). Use of the WISC-R Performance Scale and K-ABC Non-verbal Scale with deaf children in the USA and Scotland. *School Psychology International. Vol 10*(3), 193-197.

Glutting, J. J. (1986). Potthoff bias analyses of K-ABC MPC and Nonverbal Scale IQs among Anglo, Black, and Puerto Rican kindergarten children. *Professional School Psychology, 1,* 225-234.

Hernandez, A. E., & Willson, V. (1992). A comparison of Kaufman Assessment Battery for Children reliability for Mexican-Americans and non-Hispanic Whites. *Hispanic Journal of Behavioral Sciences, 14,* 394-397.

Hodapp, R. M., Leckman, J. F., Dykens, E. M., Sparrow, S. S et al. (1992). K-ABC profiles in children with fragile X syndrome, Down syndrome, and nonspecific mental retardation. *American Journal on Mental Retardation, 97,* 39-46.

Krohn, E. J., & Lamp, R. E. (1989). Concurrent validity of the Stanford-Binet Fourth Edition and K-ABC for Head Start children. *Journal of School Psychology, 27,* 59-67.

Krohn, E. J; Lamp, R. E., Phelps, C. G. (1988). Validity of the K-ABC for a Black preschool population. *Psychology in the Schools, 25,* 15-21.

Lamp, R. E., & Krohn, E. J. (1990). Stability of the Stanford-Binet Fourth Edition and K-ABC for young Black and White children from low income families. *Journal of Psychoeducational Assessment, 8,* 139-149.

Lyon, Mark A; Smith, Douglas K; Klass, Patricia D.(1988) Comparison of K-ABC performance between at-risk and normal preschool children. *Perceptual & Motor Skills. Vol 66*(2), 619-626.

Matazow, G. S., Kamphaus, R. W., Stanton, H. C., & Reynolds, C. R. (1991). Reliability of the Kaufman Assessment Battery for Children for Black and White students. *Journal of School Psychology, 29,* 37-41.

Mardell-Czudnowski, C. (1995). Performance of Asian and White children on the K-ABC: Understanding information processing differences. *Assessment, 2,* 19-29.

Morris, J. M., & Bigler, E. D. (1987). Hemispheric functioning and the Kaufman Assessment Battery for Children: Results in the neurologically impaired. *Developmental Neuropsychology, 3,* 67-79.

Nolan, R. F., Watlington, D. K., & Willson, V. L. (1989). Gifted and nongifted race and gender effects on item functioning on the Kaufman Assessment Battery for Children. *Journal of Clinical Psychology, 45,* 645-650.

Phelps, L., Leguori, S., Nisewaner, K., & Parker, M. (1993). Practical interpretations of the WISC-III with language-disordered children. In B. A. Bracken, R. S. McCallum, et al. *Wechsler Intelligence Scale for Children: Third edition. Journal of Psychoeducational Assessment. Advances in psychoeducational assessment.* (pp. 71-76). Brandon, VT, USA: Clinical Psychology Publishing Co, Inc.

Pueschel, S. M. (1988). Visual and auditory processing in children with Down syndrome. In L Nadel et al. (Eds.), *The psychobiology of Down syndrome. Issues in the biology of language and cognition.* (pp. 199-216). Cambridge, MA, USA: Mit Press. vi, 484 pp.

Ricciardi, P. W., Voelker, S., & Carter, R. A., & Shore, D. L. (1991). K-ABC sequential/simultaneous processing and language-impaired preschoolers. *Developmental Neuropsychology, 7,* 523-535.

Stavrou, E., & French, J. L. (1992). The K-ABC and cognitive processing styles in autistic children. *Journal of School Psychology, 30,* 259-267.

Ulissi, S. M., Brice, P. J., & Gibbins, S. (1989). Use of the Kaufman-Assessment Battery for Children with the hearing impaired. *American Annals of the Deaf, 134,* 283-287.

Willson, V. L., Nolan, R. F., Reynolds, C. R., Kamphaus, R. W. (1989). Race and gender effects on item functioning on the Kaufman Assessment Battery for Children. *Journal of School Psychology, 27,* 289-296.

■ Factor Structure

Elwan, F. Z. (1996). Factors structure of the Kaufman Assessment Battery for Children with Egyptian schoolchildren. *Psychological Reports, 78,* 99-110.

Bracken, B. A., Howell, K. K. (1989). K-ABC subtest specificity recalculated. *Journal of School Psychology, 27,* 335-345.

Glutting, J. J. (1986). Potthoff bias analyses of K-ABC MPC and Nonverbal Scale IQs among Anglo, Black, and Puerto Rican kindergarten children. *Professional School Psychology, 1,* 225-234.

Gridley, B. E., Miller, G., Barke, C., Fischer, W. et al. (1990). Construct validity of the K-ABC with an at-risk preschool population. *Journal of School Psychology, 28,* 39-49.

Strommen, E. (1988). Confirmatory factor analysis of the Kaufman Assessment Battery for Children: A reevaluation. *Journal of School Psychology, 26,* 13-23.

Swanson, H. L., Brandenburg-Ayres, S., & Wallace, S. (1989). Construct validity of the K-ABC with gifted children. *Journal of Special Education, 23,* 342-352.

Zucker, Steven; Riordan, Jean. (1990). One-year predictive validity of new and revised conceptual language measurement. *Journal of Psychoeducational Assessment, 8,* 4-8.

■ Psychometric Properties

Allard, A. F., & Pfohl, W. (1988). The Kaufman Assessment Battery for Children: A validity study with at-risk preschoolers. *Journal of Psychoeducational Assessment, 6,* 215-224.

Bloom, A. S., Allard, A. M., Zelko, F. A., Brill, W. J. et al. (1988). Differential validity of the K-ABC for lower functioning preschool children versus those of higher ability. **American Journal on Mental Retardation, 93,** 273-277.

Bracken, B. A., Howell, K. K., Harrison, T., E., Stanford, Lisa D. et al. (1991). Ipsative subtest pattern stability of the Bracken Basic Concept Scale and the Kaufman Assessment Battery for Children in a preschool sample. *School Psychology Review, 20,* 315-330.

Donders, J. (1992). Validity of the Kaufman Assessment Battery for Children when employed with children with traumatic brain injury. *Journal of Clinical Psychology, 48,* 225-230.

Glutting, J. J., Bear, G. G. (1989). Comparative efficacy of K-ABC subtests vs. WISC—R subtests in the differential classification of learning disabilities. *Learning Disability Quarterly. Vol 12*(4), 291-298.

Hendershott, J. L., Searight, H. R., Hatfield, J. L., & Rogers, B. J. (1990). Correlations between the Stanford-Binet, Fourth Edition and the Kaufman Assessment Battery for Children for a preschool sample. *Perceptual & Motor Skills, 71,* 819-825.

Hooper, S. R., Brown, L. A., & D'Elia, F. A. (1988). A comparison of the K-ABC with the Woodcock-Johnson Tests of Academic Achievement in a referred population. *Journal of Psychoeducational Assessment, 6,* 67-77.

Jennings, W. B., Bennett, R., Cole, T., Gibson, K., et al. (1989). Commonality of diagnostic categories for students assessed on the K-ABC and WISC—R. *Journal of Psychoeducational Assessment. 7,* 74-82.

Keith, T. Z., & Novak, C. G. (1987). Joint factor structure of the WISC—R and K-ABC for referred school children. *Journal of Psychoeducational Assessment, 5,* 370-386.

Kennedy, M. H., & Hiltonsmith, R. W. (1988). Relationship among the K-ABC Nonverbal Scale, the

Pictorial Test of Intelligence, and the Hiskey-Nebraska Test of Learning Aptitude for speech-and language-disabled preschool children. *Journal of Psychoeducational Assessment, 6,* 49-54.

Knight, B. C., Baker, E. H., & Minder, C. C. (1990). Concurrent validity of the Stanford-Binet: Fourth Edition and Kaufman Assessment Battery for Children with learning-disabled students. *Psychology in the Schools, 27,* 116-125.

Krohn, E. J., & Lamp, R. E. (1989). Concurrent validity of the Stanford-Binet Fourth Edition and K-ABC for Head Start children. *Journal of School Psychology, 27,* 59-67.

Kutsick, K. A, & Wynn, E. E. (1988). Comparison of the K—ABC Achievement Scale and WPPSI IQs of preschool children. *Psychological Reports, 63,* 143-146.

Lamp, R. E., & Krohn, E., J. (1990). Stability of the Stanford-Binet Fourth Edition and K-ABC for young Black and White children from low income families. *Journal of Psychoeducational Assessment, 8,* 139-149.

Matazow, G. S., Kamphaus, R. W., Stanton, H. C., & Reynolds, C. R. (1991). Reliability of the Kaufman Assessment Battery for Children for Black and White students. *Journal of School Psychology, 29,* 37-41.

Moon, S., Ishikuma, T., & Kaufman, A. S. (1987). Joint factor analysis of the K-ABC and McCarthy scales. *Perceptual & Motor Skills, 65,* 699-704.

Smith, D. K., Bolin, J. A., & Stovall, D. L. (1988). K-ABC stability in a preschool sample: A longitudinal study. *Journal of Psychoeducational Assessment, 6,* 396-403.

Smith, D. K., Lyon, M. A, Hunter, E., & Boyd, R. (1988). Relationship between the K-ABC and WISC—R for students referred for severe learning disabilities. *Journal of Learning Disabilities, 21,* 509-513.

Rothlisberg, B. A., & McIntosh, D. E. (1991). Performance of a referred sample on the Stanford-Binet IV and the K-ABC. *Journal of School Psychology, 29,* 367-370.

Rust, J. O., & Yates, A. G. (1997). Concurrent validity of the Wechsler Intelligence Scale for Children—Third edition and the Kaufman Assessment Battery for Children. *Psychological Reports, 80,* 89-90.

Ulissi, S. M., Brice, P. J., & Gibbins, S. (1989). Use of the Kaufman-Assessment Battery for Children with the hearing impaired. *American Annals of the Deaf, 134,* 283-287.

Whitworth, R. H., & Chrisman, S. M. (1987). Validation of the Kaufman Assessment Battery for Children comparing Anglo and Mexican-American preschoolers. *Educational & Psychological Measurement, 47,* 695-702.

Williams, J. M., Voelker, S., & Ricciardi, P. W. (1995). Predictive validity of the K-ABC for exceptional preschoolers. *Psychology in the Schools, 32,* 178-185.

Zucker, S., & Copeland, E. P. (1988). K-ABC and McCarthy scale performance among "at-risk" and normal preschoolers. *Psychology in the Schools, 25,* 5-10.

■ Translations

Bartmann, U., & Kiese-Himmel, C. (1996). Die Vergleichbarkeit zweier Me_Smethoden anhand einer Studie zur Untersuchung des Wortschatzes bei sprachentwicklungsgestoerten Kindern. *Zeitschrift Fuer Differentielle und Diagnostische Psychologie, 17,* 56-61.

Berg, M., & Melchers, P. (1997). Testrezension zu Kaufman-Assessment Battery for Children (K-ABC). *Zeitschrift Fuer Differentielle und Diagnostische Psychologie, 18,* 20-24.

Giordani, B., Boivin, M. J., Opel, B., Dia N., Diawaku N., & Lauer, R. E. (1996). Use of the K-ABC with children in Zaire, Africa: An evaluation of the sequential-simultaneous processing distinction within an intercultural context. *International Journal of Disability, Development & Education, 43,* 5-24.

Gregoire, J. (1995). Application de la methode de Mantel-Haenszel a l'analyse du fonctionnement differentiel des items du K-ABC entre filles et garcons. *European Review of Applied Psychology/ Revue Europeenne de Psychologie Appliquee, 45,* 111-119.

Gregoire, J. (1995). La Kaufman Assessment Battery for Children (K-ABC). Un progres pour l'evaluation diagnostique? *Bulletin de Psychologie Scolaire et d'Orientation, 44,* 65-85.

Kiese-Himmel, C. (1995). Aktive Wortschatztestung im fruehen Kindesalter—ein Methodenvergleich bei sprachentwicklungsrueckstaendigen Kindern. *Diagnostica, 41,* 189-202.

Kiese-Himmel, C., & Kruse, E. (1995). Expressiver Wortschatz: Vergleich zweier psychologischer Testverfahren bei Kindergartenkindern. *Praxis der Kinderpsychologie und Kinderpsychiatrie, 44,* 44-47.

Melchers, P., & Preuss, U. (1992), Bearbeitung der Kaufman-Assessment Battery for Children fuer den deutschsprachigen Raum. Teil 1: Vorstellung des Verfahrens. *Zeitschrift Fuer Kinder- und Jugendpsychiatrie, 20,* 85-93.

Melchers, P., & Preuss, U. (1992). Bearbeitung der Kaufman-Assessment Battery for Children (K-ABC) fuer den deutschsprachigen Raum. Teil 2: Anwendungsbereich und Guetekriterien. *Zeitschrift Fuer Kinder- und Jugendpsychiatrie, 20,* 223-231.

Stassen, M. (1993). Un nouveau test individuel d'intelligence: le K-ABC. *Bulletin de Psychologie Scolaire et d'Orientation, 42,* 169-175.

Suess-Burghart, H. (1994). A validation study of the "Kaufman Assessment Battery for Children (K-ABC)" and the "Hamburger-Wechsler Intelligenztest fuer Kinder (HAWIK)". *Zeitschrift Fuer Differentielle und Diagnostische Psychologie, 15,* 41-47.

The Woodcock-Johnson Psycho-Educational Battery Revised (WJ-R)

■ Test Description, Administration, and Interpretation

Evans, L. D., Tannehill, R., & Martin, S. (1995). Children's reading skills: A comparison of traditional and computerized assessment. Behavior Research Methods, *Instruments, & Computers, 27,* 162-165.

Hessler, G. L. (1993). Clinical use of the Woodcock-Johnson Psycho-Educational Battery—Revised for the identification and instructional programing of types of learning disorders. In B. A. Bracken, R. S. McCallum et al. *Woodcock-Johnson Psycho-Educational Battery—Revised. Journal of Psychoeducational Assessment. Advances in psychoeducational assessment.* (pp. 123-135). Brandon, VT: Clinical Psychology Publishing Co, Inc.

McGhee, R., L., & Buckhalt, J. A. (1993). Test review. In B. A. Bracken, R. S. McCallum et al. *Woodcock-Johnson Psycho-Educational Battery—Revised. Journal of Psychoeducational Assessment. Advances in psychoeducational assessment.* (pp. 136-149). Brandon, VT: Clinical Psychology Publishing Co, Inc.

McGrew, K. S. (1993). The relationship between the Woodcock-Johnson Psycho-Educational Battery—Revised Gf-Gc cognitive clusters and reading achievement across the life-span. In B. A. Bracken, R. S. McCallum et al. *Woodcock-Johnson Psycho-Educational Battery—Revised. Journal of Psychoeducational Assessment. Advances in psychoeducational assessment.* (pp. 39-53). Brandon, VT: Clinical Psychology Publishing Co, Inc.

McGrew, K. S., Murphy, S. R., & Knutson, D. J. (1994). The development and investigation of a graphic scoring system for obtaining derived scores for the WJ—R and other tests. *Journal of Psychoeducational Assessment, 12,* 33-41.

Schrank, F. A. (1993). Unique contributions of the Woodcock-Johnson Psycho-Educational Battery—Revised to psychoeducational assessment. In B. A. Bracken, R. S. McCallum et al. *Woodcock-Johnson Psycho-Educational Battery—Revised. Journal of Psychoeducational Assessment. Advances in psychoeducational assessment.* (pp. 71-79). Brandon, VT: Clinical Psychology Publishing Co, Inc.

■ Special populations

Mather, N. (1993). Critical issues in the diagnosis of learning disabilities addressed by the Woodcock-Johnson Psycho-Educational Battery—Revised. In B. A. Bracken, R. S. McCallum et al. *Woodcock-Johnson Psycho-Educational Battery—Revised. Journal of Psychoeducational Assessment. Advances in psychoeducational assessment.* (pp. 103-122). Brandon, VT: Clinical Psychology Publishing Co, Inc.

■ Psychometric properties

McGrew, K. & Murphy, S. (1995). Uniqueness and general factor characteristics of the Woodcock-Johnson Tests of Cognitive Ability—Revised. *Journal of School Psychology, 33,* 235-245.

McGrew, K. S. & Hessler, G. L. (1995). The relationship between the WJ—R Gf-Gc cognitive clusters and mathematics achievement across the life-span. *Journal of Psychoeducational Assessment, 13,* 21-38.

McGrew, K. S., & Knopik, S. N. (1993). The relationship between the WJ—R Gf-Gc cognitive clusters and writing achievement across the life-span. *School Psychology Review, 22,* 687-695.

Sinnett, E. R., & Rogg, K. L., Benton, S. L., Downey, R., G. et al. (1993). The Woodcock-Johnson Revised: Its factor structure. *Educational & Psychological Measurement, 53,* 763-769.

Wechsler Intelligence Scales for Children and Wechsler Preschool and Primary Scales of Intelligence (WISC and WPPSI)

■ Descriptions, Administration, and Interpretation

Burgess, A. (1991). Profile analysis of the Wechsler intelligence scales: A new index of subtest scatter. *British Journal of Clinical Psychology, 30,* 257-263.

Glutting, J. J., & McDermott, P. A. (1990). Patterns and prevalence of core profile types in the WPPSI standardization sample. *School Psychology Review, 19,* 471-491.

Gyurke, J. S., Prifitera, A., & Sharp, S. A. (1991). Frequency of verbal and performance IQ discrepancies on the WPPSI—R at various levels of ability. *Journal of Psychoeducational Assessment, 9,* 230-239.

Kaplan, C. H., Fox, L. M., & Paxton, L. (1991). Bright children and the revised WPPSI: Concurrent validity. *Journal of Psychoeducational Assessment, 9,* 240-246.

Kaplan, C. (1992). Ceiling effects in assessing high-IQ children with the WPPSI—R. *Journal of Clinical Child Psychology, 21,* 403-406.

Kaufman, A. S. (1992). Evaluation of the WISC-III and WPPSI—R for gifted children. *Roeper Review, 14,* 154-158.

Kaufman, A. S. (1990). The WPPSI—R: You can't judge a test by its colors. *Journal of School Psychology, 28,* 387-394.

Keith, T. Z. (1994). Intelligence is important, intelligence is complex. *School Psychology Quarterly, 9,* 209-221.

LoBello, S. G. (1991). A short form of the Wechsler Preschool and Primary Scale of Intelligence—Revised. *Journal of School Psychology, 29,* 229-236.

LoBello, S. G. (1991). Subtest scatter as an indicator of the inaccuracy of short-form estimates of IQ. *Psychological Reports, 68,* 1115-1118.

LoBello, S. G. (1991). A table for determining probability of obtaining verbal and performance scale discrepancies on the Wechsler Preschool and Primary Scale of Intelligence—Revised. *Psychology in the Schools, 28,* 93-94.

LoBello, S. G. (1991). Significant differences between individual subtest scaled scores and average scaled scores on the Wechsler Preschool and Primary Scale of Intelligence—Revised. *Psychology in the Schools. 28,* 15-18.

Milrod, R. J., & Rescorla, L. (1991). A comparison of the WPPSI—R and WPPSI with high-IQ children. *Journal of Psychoeducational Assessment, 9,* 255-262.

Novak, P. A., Tsushima, W. T., & Tsushima, M. M. (1991). Predictive validity of two short-forms of the WPPSI: A 3-year follow-up study. *Journal of Clinical Psychology, 47,* 698-702.

Quereshi, M. Y., & Seitz, R. (1994). Non-equivalence of WPPSI, WPPSI—R, and WISC—R scores. *Current Psychology: Developmental, Learning, Personality, Social, 13,* 210-225.

Quereshi, M. Y., & Seitz, R. (1994). Gender differences on the WPPSI, the WISC—R, and the WPPSI—R. *Current Psychology: Developmental, Learning, Personality, Social, 13,* 117-123.

Razavieh, A., & Shahim, S. (1992). A short form of the Wechsler Preschool and Primary Scale of Intelligence for use in Iran. Psychological Reports, 71, 863-866.

Roid, G. H., & Gyurke, J. (1991). General-factor and specific variance in the WPPSI—R. *Journal of Psychoeducational Assessment, 9,* 209-223.

Sattler, J. M. (1992). Assessment of children: WISC—III and WPPSI—R supplement. San Diego, CA: Jerome M. Sattler.

Sattler, J. M. (1991). Normative changes on the Wechsler Preschool and Primary Scale of Intelligence—Revised Animal Pegs subtest. *Psychological Assessment, 3,* 691-692.

Silverstein, A. B. (1987). Unusual combinations of Verbal and Performance IQs on Wechsler's intelligence scales. *Journal of Clinical Psychology, 43,* 720-722.

Silverstein, A. B. (1987). Two indices of subtest scatter on Wechsler's intelligence scales: Estimated vs. empirical values. *Journal of Clinical Psychology, 43,* 409-414.

Silverstein, A. B. (1990). Notes on the reliability of Wechsler short forms. *Journal of Clinical Psychology, 46,* 194-196.

Speer, S. K., Hawthorne, Linda W., & Buccellato, L. (1986). Intellectual patterns of young gifted children on the WPPSI. *Journal for the Education of the Gifted, 10,* 57-62.

Towle, P (1989). The Wechsler Preschool and Primary Scales of Intelligence. In C. S. Newmark, (Ed), et al. (1989). *Major psychological assessment instruments, Vol. 2.* (pp. 251-270). Needham Heights, MA: Allyn & Bacon.

Tsushima, W. T. (1994). Short form of the WPPSI and WPPSI—R. *Journal of Clinical Psychology, 50,* 877-880.

Whitten, J., Slate, J. R., Jones, C. H., & Shine, A. E. (1994). Examiner errors in administering and scoring the WPPSI—R. *Journal of Psychoeducational Assessment, 12,* 49-54.

■ Special Populations

McEvoy, R. E., & Johnson, D. L. (1989). Comparison of an intelligence test and a screening battery as predictors of reading ability in low income, Mexican American children. *Hispanic Journal of Behavioral Sciences.* Vol 11(3), 274-282.

Kaufman, A. S. (1992). Evaluation of the WISC-III and WPPSI—R for gifted children. *Roeper Review, 14,* 154-158.

Keith, T. Z. (1994) Intelligence is important, intelligence is complex. *School Psychology Quarterly. Vol 9*(3), 209-221.

Reich, J. N., Cleland, J. W., Stilson, S. R., Kaspar, J. C., et al. (1993) Children born at risk: What's happening in kindergarten? *Psychology in the Schools, 30,* 50-52.

■ Psychometric Properties

Carvajal, H., Hardy, K., Smith, K. L., & Weaver, K. A. (1988). Relationships between scores on Stanford-Binet IV and Wechsler Preschool and Primary Scale of Intelligence. *Psychology in the Schools, 25,* 129-131.

Carvajal, H. H., Parks, J. P., Bays, K. J., Logan, R. A., et al. (1991). Relationships between scores on Wechsler Preschool and Primary Scale of Intelligence—Revised and Stanford-Binet IV. *Psychological Reports, 69,* 23-26.

Carvajal, H. H., Parks, J. P., Logan, R. A., & Page, G. L. (1992). Comparisons of the IQ and Vocabulary scores on Wechsler Preschool and Primary Scale of Intelligence—Revised and Peabody Picture Vocabulary Test—Revised. *Psychology in the Schools, 29,* 22-24.

Carvajal, H. H., Parks, C. S., Parks, J. P., Logan, R. A., et al. (1993). A concurrent validity study of the Wechsler Preschool and Primary Scale of Intelligence—Revised and Columbia Mental Maturity Scale. *Bulletin of the Psychonomic Society, 31,* 33-34.

Chermak, G. D., & Fisher, J. M. (1989). Association between paired subtests of auditory sequential memory administered to preschool children. *Perceptual & Motor Skills, 68,* 255-258.

Corkum, V., & Dunham, P. (1996). The Communicative Development Inventory-WORDS Short Form as an index of language production. *Journal of Child Language, 23,* 515-528.

Faust, D. S., & Hollingsworth, J. O. (1991). Concurrent validation of the Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI—R) with two criteria of cognitive abilities. *Journal of Psychoeducational Assessment, 9,* 224-229.

Gerken, K. C., & Hodapp, A. F. (1992). Assessment of preschoolers at-risk with the WPPSI—R and the Stanford-Binet L-M. *Psychological Reports, 71,* 659-664.

Glutting, J. J., & McDermott, P. A. (1989). Using "teaching items" on ability tests: A nice idea, but does it work? *Educational & Psychological Measurement, 49,* 257-268.

Gyurke, J. S., Stone, B. J., & Beyer, M. (1990). A confirmatory factor analysis of the WPPSI—R. *Journal of Psychoeducational Assessment, 8,* 15-21.

Kaplan, C. (1993). Predicting first-grade achievement from pre-kindergarten WPPSI—R scores. *Journal of Psychoeducational Assessment, 11,* 133-138.

Kaplan, C. (1993). Reliability and validity of test-session behavior observations: Putting the horse before the cart. *Journal of Psychoeducational Assessment, 11,* 314-322.

Kaplan, C. (1996). Predictive validity of the WPPSI—R: A four year follow-up study. *Psychology in the Schools, 33,* 211-220.

Karr, S. K., Carvajal, H. H., Elser, D., Bays, K., et al. (1993). Concurrent validity of the WPPSI—R and the McCarthy Scales of Children's Abilities. *Psychological Reports, 72,* 940-942.

Kutsick, K., Vance, B., Schwarting, F. G., & West, R. (1988). A comparison of three different measures of intelligence with preschool children identified at-risk. *Psychology in the Schools, 25,* 270-275.

Kutsick, K. A., & Wynn, E. E. (1988). Comparison of the K—ABC Achievement Scale and WPPSI IQs of preschool children. *Psychological Reports, 63,* 143-146.

Lewis, C. D., & Lorentz, S. (1994). Comparison of the Leiter International Performance Scale and the Wechsler Intelligence Scales. *Psychological Reports, 74,* 521-522.

Laughlin, T. (1995). The school readiness composite of the Bracken Basic Concept Scale as an intellectual screening instrument. *Journal of Psychoeducational Assessment, 13,* 294-302.

LoBello, S. G., & Guelgoez, S. (1991). Factor analysis of the Wechsler Preschool and Primary Scale of Intelligence-Revised. *Psychological Assessment, 3,* 130-132.

Macmann, G. M., & Barnett, D. W. (1994). Structural analysis of correlated factors: Lessons from the verbal performance dichotomy of the Wechsler scales. *School Psychology Quarterly, 9,* 161-197.

McEvoy, R. E., & Johnson, D. L. (1989). Comparison of an intelligence test and a screening battery as predictors of reading ability in low income, Mexican American children. *Hispanic Journal of Behavioral Sciences, 11,* 274-282.

Moore, C., O'Keefe, S. L., & Lawhon, D. (1998). Concurrent validity of the Snijders-Oomen Nonverbal Intelligence Test 2 1/2-7—Revised with the Wechsler Preschool and Primary Scale of Intelligence—Revised. *Psychological Reports, 82,* 619-625.

Neyens, L. G. J, & Aldenkamp, A. P. (1997). Stability of cognitive measures in children of average ability. *Child Neuropsychology, 3,* 161-170.

Novak, P. A., Tsushima, W. T., & Tsushima, M. M. (1991). Predictive validity of two short-forms of the WPPSI: A 3-year follow-up study. *Journal of Clinical Psychology, 47,* 698-702.

O'Grady, K. E. (1990). A confirmatory maximum likelihood factor analysis of the WPPSI. *Personality & Individual Differences, 11,* 135-140.

Quereshi, M. Y., & Seitz, R. (1994). Non-equivalence of WPPSI, WPPSI—R, and WISC—R scores. *Current Psychology: Developmental, Learning, Personality, Social, 13,* 210-225.

Quereshi, M. Y., & Seitz, R. (1994). Gender differences on the WPPSI, the WISC—R, and the WPPSI—R. *Current Psychology: Developmental, Learning, Personality, Social, 13,* 117-123.

Razavieh, A., & Shahim, S. (1990) Retest reliability of the Wechsler Preschool and Primary Scale of Intelligence restandardized in Iran. *Psychological Reports, 66,* 865-866.

Roid, G. H., & Gyurke, J. (1991). General-factor and specific variance in the WPPSI—R. *Journal of Psychoeducational Assessment, 9,* 209-223.

Sabers, D., Jones, P., & Shirome, P. (1989). On methods for probing validity of intelligence tests: A commentary on the work of Zeidner and Feitelson. *Journal of Psychoeducational Assessment, 7,* 194-208.

Sattler, J. M., & Atkinson, L. (1993). Item equivalence across scales: The WPPSI—R and WISC-III. *Psychological Assessment, 5,* 203-206.

Schneider, B. H., & Gervais, M. D. (1991). Identifying gifted kindergarten students with brief screening measures and the WPPSI—R. *Journal of Psychoeducational Assessment, 9,* 201-208.

Silverstein, A. B. (1991). Reliability of score differences on Wechsler's intelligence scales. *Journal of Clinical Psychology, 47,* 264-266.

Stone, B. J., Gridley, B. E., & Gyurke, J. S. (1991). Confirmatory factor analysis of the WPPSI—R at the extreme end of the age range. *Journal of Psychoeducational Assessment, 9,* 263-270.

Urbina, S. P., Clayton, J. P. (1991). WPPSI—R/WISC—R: A comparative study. *Journal of Psychoeducational Assessment, 9,* Sep, 247-254.

Vance, B., West, R., & Kutsick, K. (1989). Prediction of Wechsler Preschool and Primary Scale of Intelligence IQ scores for preschool children using the Peabody Picture Vocabulary Test—R and the Expressive One Word Picture Vocabulary Test. *Journal of Clinical Psychology, 45,* 642-644.

Zeidner, M., & Feitelson, D. (1989). Probing the validity of intelligence tests for preschool children: A smallest space analysis. *Journal of Psychoeducational Assessment, 7,* 175-193.

■ Translations

Brzezinska, A., Czub, T., Lutomski, G., & Mallecka, M. (1991). Polish adaptation of the Wechsler Preschool and Primary Scale of Intelligence (WPPSI): Preliminary report. *Psychological Bulletin, 22,* 289-301.

Ottem, E., & Sletmo, A. (1993). Speech and language impairment in preschool years and reading difficulties in school age: Predictions on the basis of WPPSI test profiles. *Tidsskrift for Norsk Psykologforening, 30,* 335-341.

Razavieh, A., & Shahim, S. (1992). A short form of the Wechsler Preschool and Primary Scale of Intelligence for use in Iran. *Psychological Reports, 71,* 863-866.

Shahim, S. (1992). Correlations for Wechsler Intelligence Scale for Children—Revised and the Wechsler Preschool and Primary Scale of Intelligence for Iranian children. *Psychological Reports, 70,* 27-30.

The Vineland Adaptive Behavior Scales

Altepeter, T., Moscato, E., & Cummings, J. (1986). Comparison of scores of hearing-impaired children on the Vineland Adaptive Behavior Scales and the Vineland Social Maturity Scale. *Psychological Reports, 59,* 635-639.

Atkinson, L. (1990). Standard errors of prediction for the Vineland Adaptive Behavior Scales. *Journal of School Psychology, 28,* 355-359.

Atkinson, L. (1990) Intellectual and adaptive functioning: Some tables for interpreting the Vineland in combination with intelligence tests. *American Journal on Mental Retardation, 95,* 198-203.

Atkinson, L., Bevc, I., Dickens, S., & Blackwell, J. (1992). Concurrent validities of the Stanford-Binet (Fourth Edition), Leiter, and Vineland with developmentally delayed children. *Journal of School Psychology, 30,* 165-173.

Britton, W., & Eaves, R. (1986). Relationship between the Vineland Adaptive Behavior Scales—classroom edition and the Vineland Social Maturity Scales. *American Journal of Mental Deficiency, 91,* 105-107.

Cohen, H. (1988). Measurement of adaptive behavior: Origins, trends, issues. *Child & Youth Services, 10,* 37-81.

de Lemos, M. (1989). The Vineland Adaptive Behaviour scales: standard score adjustments for Australian children. *Psychological Test Bulletin, 2,* 3-15.

Doll, E. (1988). Before the big time: Early history of the training school at Vineland, 1888 to 1949. *American Journal on Mental Retardation, 93,* 1-15.

Douthitt, V. (1992). A comparison of adaptive behavior in gifted and nongifted children. *Roeper Review, 14,* 149-151.

Eggert, D. (1974). A comparative study on the social competence of trainable mentally retarded children and non-retarded children with the Vineland Social Maturity Scale. *Praxis der Kinderpsychologie und Kinderpsychiatrie, 23,* 139-144.

Evans, M., & Wodar, S. (1997). Maternal sensitivity to vocabulary development in specific language-impaired and language-normal preschoolers. *Applied Psycholinguistics, 18,* 243-256.

Fromme, D. (1974). On the use of the Vineland Social Maturity Scale as an estimate of intellectual functioning. *Journal of Clinical Psychology 30,* 67-68.

Givens, T. (1978). The current status of three major techniques for the assessment of social competence in the diagnosis of the potentially retarded child. *Southern Journal of Educational Research, 12,* 75-84.

Greensen, M. (1974). *Retarded children: An analysis of test scores.* Tilburg, Netherlands: Tilburg University Press.

Hooshyar, N. (1987). Relationship between maternal language parameters and the child's language competency and developmental condition. *International Journal of Rehabilitation Research, 10,* 321-324.

Hundert, J., Morrison, L., Mahoney, W., Mundy, F., & Vernon, M. (1997). Parent and teacher assessments of the developmental status of children with severe, mild/moderate, or no developmental disabilities. *Topics in Early Childhood Special Education, 17,* 419-434.

Jenkins, J., & Guidubaldi, J. (1997). The nature-nurture controversy revisited: Divorce and gender as factors in children's racial group differences. *Child Study Journal, 27,* 145-160.

Johnson, L., Cook, M., & Kullman, A. (1992). An examination of the concurrent validity of the Battelle Developmental Inventory as compared with the Vineland Adaptive Scales and the Bayley Scales of Infant Development. *Journal of Early Intervention, 16,* 353-359.

Kaplan, H., & Alatishe, M. (1976). Comparison of ratings by mothers and teachers on preschool children using the Vineland Social Maturity Scale. *Psychology in the Schools, 13,* 27-28.

Loveland, K., & Kelley, M. (1988). Development of adaptive behavior in adolescents and young adults with autism and Down syndrome. *American Journal on Mental Retardation, 93,* 84-92.

Middleton, H., Keene, R., & Brown, G. (1990). Convergent and discriminant validities of the scales of independent behavior and the revised Vineland Adaptive Behavior Scales. *American Journal on Mental Retardation, 94,* 669-673.

Pearson, D., & Lachar, D. (1994). Using behavioral questionnaires to identify adaptive deficits in elementary school children. *Journal of School Psychology, 32,* 33-52.

Pedrini, D. Pedrini, B., & Gregory, L. (1973). Aids in the administration, scoring, or interpretation of the Stanford-Binet, Vineland, and Wechsler Scales. *Catalog of Selected Documents in Psychology, 3,* 128.

Perry, A., & Factor, D. (1989). Psychometric validity and clinical usefulness of the Vineland Adaptive Behavior Scales and the AAMD Adaptive Behavior Scale for an autistic sample. *Journal of Autism & DevelopmentalDisorders, 19,* 41-55.

Poth, R., & Barnett, D. (1988). Establishing the limits of interpretive confidence: A validity study of two preschool developmental scales. *School Psychology Review, 17,* 322-330.

Quarrington, B., & Solomon, B. (1975). A current study of the social maturity of deaf students. *Canadian Journal of Behavioural Science, 7,* 70-77.

Quinn, J. (1999). Transcultural assessment: Toward competence for the twenty-first century. In J. McFadden (Ed.), *Transcultural counseling.* (2nd edition) (pp. 343-372). Alexandria: American Counseling Association.

Raggio. D., & Massingale, T. (1990). Comparability of the Vineland Social Maturity Scale and the Vineland Adaptive Behavior Scale-Survey form with infants evaluated for developmental delay. *Perceptual and Motor Skills, 71,* 415-418.

Raggio, D., & Aldridge, J. (1993). Comparison of the Vineland Social Maturity Scale, the Vineland Adaptive Behavior Scale- Survey form, and the Bayley scales of infant development with infants evaluated for developmental delay. *Perceptual and Motor Skills, 77,* 931-937.

Raggio, D., Massingale, T., & Bass, J. (1994). Comparison of the Vineland Adaptive Behavior Scales-Survey form age equivalent and standard score with the Bayley mental development index. *Perceptual and Motor Skills, 79,* 203-206.

Roszkowski, M. (1980). Concurrent validity of the Adaptive Behavior Scale as assessed by the the Vineland Social Maturity Scale. *American Journal of Mental Deficiency 85,* 86-89.

Schatz, J., & Hamdan-Allen, G. (1995). Effects of age and IQ on adaptive behavior domains for children with autism. *Journal of Autism & Developmental Disorders, 25,* 51-60.

Simon, E., Rosen, M., Grossman, E., & Pratowski, E. (1995). The relationships among facial emotion recognition, social skills, and quality of life. *Research in Developmental Disabilities, 16,* 383-391

Song, A., & Jones, S. (1982). Vineland Social Maturity Scale norm examined: The Wisconsin experience with 0-to 3-year-old children. *American Journal of Mental Deficiency, 86,* 428-431.

Sparrow, S. (1984). *Vineland Adaptive Behavior Scales: interview edition, survey form manual.* Circle Pines, MN: American Guidance Service.

Szatmari, P., Archer, L., Fisman, S., & Steiner, D. (1994). Parent and teacher agreement in the assessment of pervasive developmental disorders. *Journal of Autism & Developmental Disorders, 24,* 703-717.

Tombokan-Runtukahu, J., & Nikto, A. (1992). Translation, cultural adjustment, and validation of a measure of adaptive behavior. *Research in Developmental Disabilities, 13,* 481-501.

Tucker, C., Brady, B., Harris, Y., Fraser, K., et al. (1993). The association of selected parent behaviors with the adaptive and maladaptive functioning of Black children and White children. *Child Study Journal, 23,* 39-55.

VanMeter, L., Fein, D., Morris, R., Waterhouse, L., & Allen, D. (1997). Delay versus deviance in autistic social behavior. Journal of Autism & Developmental Disorders, 27, 557-569.

Vig, S., & Jedrysek, E. (1995). Adaptive behavior of young urban children with developmental disabilities. *Mental Retardation, 33,* 90-98.

Voelker, S., Shore, D., & Miller, L. (1987). Vineland Adaptive Behavior Scales with mentally retarded adults: Informant versus self-report. *Mental Retardation & Learning Disability Bulletin, 15,* 21-28.

Voelker, S., Shore, D., Brown-More, C., Hill, L. et al. (1990) Validity of self-report of adaptive behavior skills by adults with mental retardation. *Mental Retardation, 28,* 305-309.

Voelker, S., Shore, D., Hakim-Larson, J., & Bruner, D. (1997). Discrepancies in parent and teacher ratings of adaptive behavior of children with multiple disabilities. *Mental Retardation, 35,* 10-17.

Wodrich, D., & Barry, C. (1991). A survey of school psychologists' practices for identifying mentally retarded students. *Psychology in the Schools, 28,* 165-171.

The Raven Progressive Matrices

Fields, J. (1997). Measuring giftedness in young children: a comparative study in Malaysia. *Early Child Development and Care, 131,* 93-106.

Gudjonsson, G. (1995). The Standard Progressive Matrices: methodological problems associated with the administration of the 1992 adult standardization sample. *Personality and Individual Differences, 18,* 441-442.

Gudjonsson, G. (1995). Raven's norms on the SPM revisited: a reply to Raven. *Personality and Individual Differences, 18,* 447.

MacAvoy, J., Orr, S., & Sidles, C. (1993). The Raven Matrices and Navajo children: normative characteristics and culture fair application to issues of intelligence, giftedness, and academic proficiency. *Journal of American Indian Education, Fall,* 33-43.

Morris, G., & Alcorn, M. (1995). Raven's Progressive Matrices and inspection time: P200 slope correlates. *Personality and Individual Difference, 18,* 81-87.

Powers, S., & Barkan, J. (1986). Concurrent validity of the standard progressive matrices for Hispanic and NonHispanic seventh-grade students. *Psychology in the Schools, 23,* 333-336.

Wright, S., Taylor, D., & Ruggiero, K. (1996). Examining the potential for academic achievement among Inuit children. Comparisons on the Raven Colored Progressive Matrices. *Journal of Cross-Cultural Psychology, 27,* 733-753.

EARLY CHILD
DEVELOPMENT

EDUCATION
THE WORLD BANK
HUMAN DEVELOPMENT NETWORK

For more copies, contact
Mary Eming Young
The World Bank
1818 H Street, N.W.
Washington, D.C. 20433
Phone: (202) 473-3427
Fax: (202) 522- 3233
Email: myoung3@worldbank.org

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").