Table of Contents

[**] Question as to whether or not the math-related subtest can be used independently.

iv. Devereaux Early Childhood Assessment (DECA)
v. Social Competence and Behavior Evaluation—Preschool Edition (SCBE)
vi. Social Skills Rating System (SSRS)
vii. Social Skills Rating System (SSRS), Task Orientation/Approaches to Learning Scale
viii. Vineland Social-Emotional Early Childhood Scales (SEEC)
ix. Social-Emotional/Approaches to Learning Reference List

e. **Ongoing Observational Assessments:**
i. Creative Curriculum Developmental Continuum for Ages 3-5 (Creative)
ii. High/Scope Child Observation Record (COR)
iii. The Galileo System for the Electronic Management of Learning (Galileo)
iv. The Work Sampling System (WSS)
v. Ongoing Observational Assessments Reference List

II. Measures used in Early Head Start Evaluation

Authors: Allison Sidle Fuligni and Christy Brady-Smith

Center for Children and Families, Teachers College, Columbia University

i. Child-Parent Interaction Rating Scales for the Three-Bag Assessment
ii. Child-Parent Rating Scales for the Puzzle Challenge Task
iii. Nursing Child Assessment Satellite Training (NCAST): Teaching Task Scales

III. Measures used in Head Start FACES Study

Authors: Alberto Sorongon, Kwang Kim, and Nicholas Zill

WESTAT

i. Pending author permission

Early Childhood Measures: Cognitive

Early Childhood Measures: Cognitive

Bracken Basic Concept Scale—Revised (BBCS-R)

I. Background Information

Author/Source

Source: Bracken, B. A. (1998). *Bracken Basic Concept Scale—Revised: Examiner's Manual*. San Antonio, TX: The Psychological Corporation.

Publisher: The Psychological Corporation
19500 Bulverde Rd.
San Antonio, TX 78259
Phone: 800-872-1726
Website: www.psychcorp.com

Purpose of Measure

As described by instrument publisher:
This measure is designed to assess children's concept development and to determine how familiar children are with concepts that parents, preschool teachers, and kindergarten teachers teach children to prepare them for formal education.

"The BBCS-R English edition serves five basic assessment purposes: speech-language assessment, cognitive assessment, curriculum-based assessment, school readiness screening, and assessment for clinical and educational research" (Bracken, 1998, p. 6).

Population Measure Developed With

- The standardization sample was representative of the general U.S. population of children ages 2 years, 6 months through 8 years and was stratified by age, gender, race/ethnicity, region, and parent education. Demographic percentages were based on 1995 U.S. Census data.
- The sample consisted of 1,100 children between the ages of 2 years, 6 months and 8 years.
- In addition to the main sample, two clinical studies were conducted—one with 36 children who were developmentally delayed, and one with 37 children who had language disorders.

Age Range Intended For

2 years, 6 months through 8 years.

Key Constructs of Measure

The BBCS-R includes a total of 308 items in 11 subtests tapping "…foundational and functionally relevant educational concepts…" (Bracken, 1998, p. 13). The 11 subtests are as follows:

- *Colors.* Identification of primary colors and basic color terms.
- *Letters.* Knowledge of upper and lower case letters.

- *Numbers/Counting.* Number recognition and counting abilities.
- *Sizes*: Understanding of one-, two-, and three-dimensional size concepts such as tall, short, and thick.
- *Comparisons*: Matching or differentiating objects based on salient characteristics.
- *Shapes*. Knowledge of basic one-, two-, and three-dimensional shapes (e.g., line, square, cube), and abstract shape-related concepts (e.g., space).
- *Direction/Position:* Understanding of concepts such as behind, on, closed, left/right, and center.
- *Self-/Social Awareness:* Understanding of emotions such as angry and tired; understanding of terms describing kinship, gender, relative ages, and social appropriateness.
- *Texture/Material*: Understanding of terms describing characteristics of an object, such as heavy, and sharp; knowledge of composition of objects, such as wood and glass.
- *Quantity:* Understanding of concepts involving relative quantities, such as a lot, full, and triple.
- *Time/Sequence:* Understanding of concepts related to timing, duration, and ordering of events, such as after, summer, and slow (Bracken, 1998, p. 13).

A School Readiness Composite (SRC) is constructed from the first six subtests (Colors, Letters, Numbers/Counting, Sizes, Comparisons, and Shapes). A full battery score can also be created using all 11 subtests. This review focuses on the SRC and the full battery, both of which are options for overall measures of children's concept development.

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

Comments

- The BBCS-R is different from a number of the other assessments that we have reviewed that focus on underlying ability or IQ. This measure is achievement-oriented, focusing on constructs that children learn through instruction.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
The BBCS-R is designed to minimize verbal responses. Responses are either pointing responses (i.e., the child is asked to respond by pointing to pictures) or short verbal responses. Example: "Look at all of the pictures. Show me the circle."

BBCS-R utilizes basals and ceilings. A ceiling is established within each subtest when the child answers three consecutive items incorrectly. For the first six subtests (SRC), assessment always

starts with the first item. The starting point for the rest of the subtests is determined based on the child's SRC score, and a basal is established when the child passes three consecutive items.

Who Administers Measure/Training Required?

Test Administration:

- Those who administer and interpret the results of BBCS-R should be knowledgeable in the administration and interpretation of assessments. According to the publisher, people who are involved with psychoeducational assessment or screening (school psychologists, special education teachers, etc.) will find the test easy to administer, score, and interpret.

Data Interpretation:

- (Same as above.)

Setting (e.g. one-on-one, group, etc.)

One-on-one.

Time Needed and Cost

Time:

- The BBCS-R is untimed, so the time needed for each subtest and the full battery is variable. According to Psychological Corporation's customer service, it takes about 30 minutes to administer the six subtests required to construct the SRC composite.

Cost:

- Complete kit: $245
- Examiner's Manual: $63

Comments

- As noted by the publisher, because the BBCS-R minimizes verbal responses, it can be used as a warm-up for other assessments. In addition, it is useful for children who are shy or hesitant, or for those with a variety of conditions that might limit participation in other assessments (e.g., social phobia, autism).

III. Functioning of Measure

Reliability Information from Manual

Split-Half Reliability

Split-half reliability estimates were calculated by correlating total scores on odd-numbered items with total scores on even-numbered items and applying a correction formula to estimate full-test reliabilities. As in the calculations of test-retest reliability, analyses were conducted using the SRC, subtests 7 to 11, and the full battery score. The average split-half reliabilities across ages 2 years to 7 years ranged from .91 for the SRC to .98 for the Total Test, with reliability estimates increasing slightly between ages 2 and 5 (see Bracken, 1998, p. 64).

Test-Retest Reliability
A subsample of 114 children drawn from the standardization sample took the BBCS-R twice (7 to 14 days apart). The sample was drawn from three age groups—3 years, 5 years, and 7 years. As with split-half reliability analyses, the authors did not look at tests 1 through 6 (i.e., Colors, Letters, Numbers/Counting, Sizes, Comparisons, and Shapes) separately, but instead looked at SRC composite scores. Analyses were conducted using the SRC and individual tests 7 to 11 (i.e., Direction/Position, Self-/Social Awareness, Texture/Material, Quantity, and Time/Sequence). The test-retest reliability of the SRC was .88. The test-retest reliabilities of subtests 7 to 11 were .78 for both Quantity and Time/Sequence, .80 for Texture/Material, and .82 for both Direction/Position and Self-/Social Awareness. Test-retest reliability of the Total Test was .94 (see Bracken, 1998, p. 67).

Validity Information from Manual
Internal Validity
Correlations were calculated for each age group (2 years to 7 years), as well as for the full sample, between scores on the SRC, subtests 7 to 11, and the full battery. Intercorrelations among the SRC and scores on subtests 7 to 11 for the full sample ranged from .58 (between Time/Sequence and the SRC) to .72 (between Self-/Social Awareness and Direction/Position). In the full sample, intercorrelations between subtests 7 to 11 and Total Test scores ranged from .79 (with Time/Sequence) to .87 (with Direction/Position). The intercorrelations between the SRC and the Total Test was .85, indicating that the subtests and the SRC were fairly consistent in their associations with Total Test scores (see Bracken, 1998, p. 75). Bracken (1998) concludes that these correlational findings "…support the claim that the subtests are all measuring part of a larger common theme (basic concepts), yet…indicate that the measures are [not] identical in what they are assessing…" (p. 74).

Concurrent Validity
A number of studies were reported in which children's scores on the BBCS-R were correlated with scores on other measures of cognitive, language, and conceptual development. Across these studies, correlations between BBCS-R scale scores and scores on other measures ranged from .34 to .89, with most falling above .70.

- Scores on the BBCS-R were correlated with scores on the Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R; Wechsler, 1989). The sample consisted of 30 5-year-olds (57 percent female, 43 percent male; 77 percent white, 17 percent black, 3 percent Hispanic, 3 percent other race/ethnicity).
  - Correlations between SRC scale scores and WPPSI-R scores ranged from .76 to .88, with the lowest correlation being with WPPSI-R Performance IQ scores, and the highest correlation being with WPPSI-R Full Scale IQ scores.
  - Correlations between the BBCS-R full battery scores and WPPSI-R scale scores ranged from .72 to .85, with the lowest being the correlation with WPPSI-R Performance IQ scores and the highest being the correlation with Full Scale IQ scores (see Bracken, 1998, p. 69).
- In another study, scores on BBCS-R were correlated with scores on the Differential Abilities Scale (DAS; Elliott, 1990). The sample consisted of 27 4-year-olds (67 percent female, 33 percent male; 89 percent white, 7 percent black, 4 percent other race/ethnicity).

- Correlations between SRC scores and DAS scale scores ranged from .69 to .79, with the lowest correlation being with DAS Verbal Cluster scores and the highest being with DAS General Conceptual Ability scores.
- Correlations between BBCS-R full battery scores and DAS scale scores ranged from .74 to .88, with the lowest being with DAS Verbal Cluster scores and the highest being with DAS General Conceptual Ability scale scores (see Bracken, 1998, p. 70).
    - BBCS-R scores were correlated with scores on the Boehm Test of Basic Concepts—Revised (Boehm-R; Boehm, 1986a) in a sample of 32 5-year-old children (50 percent female, 50 percent male; 72 percent white, 12.5 percent black, 12.5 percent Hispanic, 3 percent other race/ethnicity). The correlation was .73 between Boehm-R scores and SRC scores, and .89 between Boehm-R and BBCS-R full battery scores (see Bracken, 1998, p. 73).
    - BBCS-R scores were correlated with scores on the Boehm Test of Basic Concepts—Preschool Version (Boehm-Preschool; Boehm, 1986b) in a sample of 29 4-year-old children (52 percent female, 48 percent male; 66 percent white, 17 percent black, 10 percent Hispanic, 7 percent other race/ethnicity). The BBCS-R SRC correlated .34 with Boehm-Preschool scores, and BBCS-R full battery scores correlated .84 with Boehm-Preschool scores (see Bracken, 1998, p. 74).
    - Scores on the BBCS-R were correlated with scores on the Peabody Picture Vocabulary Test – Third Edition (PPVT-III; Dunn & Dunn, 1997) in a sample of 31 6-year-olds (36 percent female, 64 percent male; 84 percent white, 10 percent black, 6 percent Hispanic). PPVT-III scores correlated .69 with the SRC, and .79 with the BBCS-R full battery (see Bracken, 1998, p. 72).
    - BBCS-R scores were correlated with scores on the Preschool Language Scale – 3 (PLS-3; Zimmerman, Steiner, & Pond, 1992) scores in a sample of 27 3-year-old children (37 percent female, 63 percent male; 74 percent white, 11 percent black, 8 percent Hispanic, 7 percent other race/ethnicity).
- Correlations between SRC scores and PLS-3 scale scores ranged from .46 to .57, with the lowest being with PLS-3 Expressive Communication scores and the highest being with PLS-3 Total Language scores.
- Correlations between the BBCS-R full battery scores and PLS-3 scale scores ranged from .74 to .84, with the lowest being with PLS-3 Auditory Comprehension scores and the highest being the with PLS-3 Total Language scores (see Bracken, 1998, p. 72).

Predictive Validity

In a study of the predictive validity of the BBCS-R over the course of a kindergarten year, BBCS-R scores, children's chronological age, social skills, and perceptual motor skills were used to predict kindergartners' academic growth, as indicated by teachers' nominations for grade retention. Demographic information for this sample was not included in the Manual. These analyses correctly identified promotion/retention status for 71 of the 80 children in the sample. Among the variables included in this study, SRC scores and scores on subtests 7 through 11 were found to be the strongest predictors of children's academic growth (see Bracken, 1998, p. 71).

Discriminant Validity

A study was conducted with 37 3-, 4-, and 5-year-old children who were diagnosed with a language delay with a receptive component. The children were matched with 37 children from

the standardization sample (matched for age, gender, parent education level, race/ethnicity, and region). The resulting samples were 38 percent female, 62 percent male; 54 percent white, 38 percent black, 3 percent Hispanic, and 5 percent other race/ethnicity. The investigators found that BBCS-R scores correctly classified children as to the presence or absence of a language disorder 74 percent of the time (see Bracken, 1998, p. 76-77).

Another study was conducted with 36 3-, 4-, and 5-year-old children who were diagnosed with a cognitive deficit and a delay in at least one other area (social, adaptive, behavior, motor, or communication). The children were matched with 36 children from the standardization sample (matched for age, gender, parent education level, race/ethnicity, and region). The resulting samples were 42 percent female, 58 percent male; 72 percent white, and 28 percent black. The investigators found that BBCS-R scores could be used to correctly classify children as to the presence or absence of a developmental delay 76 percent of the time (see Bracken, 1998, p. 77-78).

Reliability/Validity Information from Other Studies

Since the BBCS-R is a fairly recent revision, few studies of its psychometric properties are available. However, several studies of the original BBCS have been published. For example, Laughlin (1995) examined the concurrent validity of the BBCS SRC and the WPPSI-R. The sample consisted of 83 white, suburban children ranging in age from 4 years, 7 months to 4 years, 10 months. The correlation between WPPSI-R Full Scale IQ scores and SRC scores was .77; the correlation between WPPSI-R Performance IQ scores and SRC scores was .56; and the correlation between WPPSI-R Verbal IQ scores and SRC scores was .76. Laughlin (1995) concluded that the SRC is a useful kindergarten screening instrument—especially because it takes a fraction of the time that it takes to administer the WPPSI-R. However, he advises against using it as a placement or classification instrument.

Comments

- Information provided by Bracken (1998) suggests that measures derived from the BBCS-R demonstrate good reliability. Estimates of internal consistency were high across ages 2 years through 5 years. In addition, test-retest correlations were high, indicating that children's relative scores on the BBCS-R were consistent across a short time interval (one to two weeks).
    - With respect to concurrent validity, research presented by Bracken (1998) and by Laughlin (1995) indicated that scores on the BBCS-R were highly correlations with other measures of cognitive, conceptual, and language development (WPPSI-R, DAS, Boehm-R, Boehm-Preschool, PPVT-III and PLS-3). There were two exceptions to this in the work described by Bracken (1998): First, the correlation between SRC scores and scores on the Boehm-Preschool was moderate and substantially lower than the correlation between BBCS-R full battery scores and Boehm-Preschool scores (.34 vs. .84). This difference in correlations is difficult to interpret, particularly given that the six scales that constitute the SRC are also included in the full battery. Second, there was a moderate correlation between scores on the SRC and on the PLS-3 Expressive Communication scale. On the whole, however, results reported by Bracken (1998) provide some support for the validity of the BBCS-R, particularly when full battery scores are used, as a

speech-language assessment and as a cognitive assessment—two of the five assessment purposes for which it was designed (see Bracken, 1998, p. 6).

- Information presented by Bracken (1998) regarding predictive and discriminant validity indicate that scores on the BBCS-R are associated with subsequent performance in school, and that children with known language or developmental delays differ demonstrate expectable differences in their BBCS-R performance. However, it should be noted there were also substantial percentages of children who were not correctly identified on the basis of their BBCS-R scores, and we would generally concur with the conclusions of Laughlin (1995) regarding the usefulness and limitations of this measure.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- The original version of the BBCS was used in the first wave of the NICHD Study of Early Child Care (NICHD ECCRN, 1999). The study had a sample of 1,364 families in multiple cities across the United States. Families were recruited in 1991, and data were collected when children were 6 months, 15 months, 24 months, and 36 months old, with further follow-up into middle childhood occurring now. The BBCS was administered to children at 36 months of age, and SRC scores were used in analyses. Observational ratings of child care quality were not significantly associated with SRC scores. However, children whose caregivers had higher levels of education (at least some college) and training (formal, post high school training in child development) had higher scores on the SRC than did children whose caregivers had lower levels of education and training.
- The original version of the BBCS (SRC only) was used in the Child Outcomes Study of the National Evaluation of Welfare-to-Work Strategies Two Year Follow-up (McGroder, Zaslow, Moore, & LeMenestrel, 2000). This study was an experimental evaluation, examining impacts on children of their mothers' (random) assignment to a JOBS welfare-to-work program or to a control group. Two welfare-to-work program approaches (a work-first and an education-first approach) were evaluated in each of three study sites, (Atlanta, Georgia; Grand Rapids, Michigan; and Riverside, California) for a total of six JOBS programs evaluated overall in this study. Children were between the ages of 3 years and 5 years at the time of their mothers' enrollment in the evaluation, and between the ages of 5 years and 7 years at the Two Year Follow-up. The Two Year Follow-up study found an impact on SRC scores in the work-first program in the Atlanta site, with children in the program group scoring higher on the SRC than did children in the control group. This study also examined the proportion of children in the program and control groups scoring at the high and low ends of the distribution for this measure (equivalent to the top and bottom quartiles in the standardization sample). For three of the six programs, a higher proportion of children of mothers assigned to a JOBS program scored in the top quartile, compared to children of mothers in the control group. In addition, in one of the six programs, children of mothers in the program group were less likely to score in the bottom quartile on the SRC than were children of mothers in the control group.

V. Adaptation of Measure

Spanish Version

Description of Adaptation:

A Spanish version of the BBCS-R is available. Spanish-language forms are designed to be used with the English-language stimulus manual. The Spanish version is to be used as a curriculum-based measure only because it is not a norm-referenced test. Field research was conducted with a sample of 193 Spanish-speaking children between the ages of 2 years, 6 months and 7 years, 11 months.

Early Childhood Measures: Cognitive

Early Childhood Measures: Cognitive

Bayley Scales of Infant Development—Second Edition (BSID-II), Mental Scale and Mental Development Index

I. Background Information

Author/Source

Source: Bayley, N. (1993). *Bayley Scales of Infant Development, Second Edition: Manual*. San Antonio, TX: The Psychological Corporation.

Publisher: The Psychological Corporation
19500 Bulverde Rd.
San Antonio, TX 78259
Phone: 800-872-1726
Website: www.psychcorp.com

Purpose of Measure

As described by instrument publisher:
The BSID-II is designed to assess the developmental status of infants and children. "The primary value of the test is in diagnosing developmental delay and planning intervention strategies" (Bayley, 1993, p. 1).

Population Measure Developed With

- BSID-II norms were derived from a national sample of 1,700 children recruited through daycare centers, health clinics, churches, and other settings, as well as through random telephone surveys conducted by marketing research firms in eight major cities. Only children born at 36 to 42 weeks gestation and without medical complications were included in the standardization sample.
- The sample was stratified with respect to age, gender, race/ethnicity, geographic region, and parent education (see Bayley, 1993, pp. 24-28).
  - One hundred children (50 girls and 50 boys) in each of 17 1-month age groups between 1 month old and 42 months old were selected. More age groups were sampled in the 1 to 12 month range than in the 13 to 42 month range because development is more rapid at younger ages.
  - The proportions of children from each racial/ethnic group (as classified by their parents) in the standardization sample closely approximated the proportion of infants and young children from each racial/ethnic group in the U.S. population according to 1988 Census Bureau data.
  - Children were recruited from sites across the country. The number of children selected for the sample from each of four geographic regions—North Central, Northeast, South, and West—closely approximated the proportion of infants and young children in the U.S. population living in each region.
  - Parents were asked to provide information on their own education levels. The proportions of children in the sample whose parents had 0 to 12 years of education (no high school diploma), 12 years of education (high school diploma), 13 to 15 years

of education, and 16 years or more of education closely approximated the proportions of parents of infants and young children in the U.S. population reporting each level of education.

Age Range Intended For
Ages 1 month through 3 years, 6 months.

Key Constructs of Measure
The BSID-II includes a total of three scales. The focus of this summary is the Mental Scale and the Mental Development Index.

- *Mental Scale:* Items on this scale assess memory, habituation, problem solving, early number concepts, generalization, classification, vocalizations, language, and social skills. Raw scores on the Mental Scale are typically converted to age-normed Mental Development Index (MDI) scores for interpretation of children's performance.
- *Motor Scale:* Items assess control of gross and fine motor skills (e.g., rolling, crawling, sitting, walking, jumping, imitation of hand movements, use of writing implements). Raw scores on the Motor Scale are typically converted to age-normed Psychomotor Development Index (PDI) scores for interpretation.
- *Behavior Rating Scale:* This scale is used by examiners to rate qualitative aspects of children's behavior during testing sessions (e.g., engagement with tasks, examiner, and caregiver; emotional regulation).

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

Comments

- The BSID-II covers multiple domains of development. Test items relate to language, emergent literacy, early mathematics ability, social development, and motor skills.

II. Administration of Measure

Who is the Respondent to the Measure?
Infant or very young child (1 month through 3 years 6 months).

If Child is Respondent, What is Child Asked to Do?
Item sets are administered to the child based on his/her chronological age, and there are established basal and ceiling rules. On the Mental Scale, if the child passes fewer than five items in the initial item set, he/she is then assessed using the next lower item set. This continues until the child gets five or more items in a set right. If the child passes all but two or fewer items in an initial item set, the next higher item set is administered until the child does not pass three or more items in a set.

The examiner records the child's responses to objects, requests, and the testing situation in general. For some items, the examiner elicits the child's response to a particular task. For other items, the examiner makes a note if the child performed the behavior at any point during the assessment session. Examples from the Mental Scale:

- Smiles when examiner smiles at any point during the assessment or, if the examiner has not seen the behavior, when the examiner explicitly smiles at the child in an attempt to get him/her to reciprocate.
- Habituates to rattle that the examiner shakes five times for 10 seconds, 12-15 inches from the child's head, just outside of his/her field of vision.
- Eyes follow ring in motion that the examiner moves above the child while he/she is lying down. (Several trials use different paths—horizontal, circular, etc.).
- Approaches mirror image.
- Removes lid from box after watching the examiner put a toy inside.
- Imitates a word at any point during the assessment or, if the child has not done so, when the examiner speaks single words to the child in an effort to get him/her to imitate.
- Counts to at least three when the examiner asks him/her to count.
- Understands concept of "more" (i.e., correctly says who has more blocks when the examiner has six and the child has two).

Who Administers Measure/Training Required?

Test Administration:

- Because the BSID-II is complex to administer, those who administer it should have training and experience with developmental assessments such as the BSID-II, as well as experience testing young children. Most examiners who use the BSID-II have completed graduate or professional training in assessment, although someone without such a background can be trained to administer the assessment if supervised closely.

Data Interpretation:

- BSID-II is also complex to interpret. Those who interpret the results should have training and experience with assessment and psychometrics. They should also have an understanding of the uses and limitations of BSID-II test results.

Setting (e.g. one-on-one, group, etc.):

One-on-one.

Time Needed and Cost

Time:

- Recommended times (according to the Manual) are 25 to 35 minutes for children under 15 months old and up to 60 minutes for children older than 15 months.

Cost:

- Complete kit: $950
- Manual: $80

Comments

- Test administration is flexible and takes into account the young age of the children being tested. The examiner can re-administer earlier items if the child is initially shy or reluctant. In addition, if a parent is present during administration, he/she can attempt to elicit the behavior for certain items. The examiner can also administer the test over two sessions if the child is restless or irritable.

III. Functioning of Measure

Reliability Information from Manual (Mental Development Index)

Internal Reliability

Internal reliability estimates (coefficient alphas) were computed for multiple 1-month age groups between 1 month and 42 months (100 children in each of 17 age groups). The average alpha for the Mental Scale was .88, ranging from .78 to .93 (see Bayley, 1993, p. 191).

Test-Retest Stability

A sample of 175 children, drawn from four age groups in the standardization sample (1, 12, 24, and 36 months), were tested twice. Children were re-tested between 1 and 16 days after their first assessment, with a median interval of 4 days. For the MDI, for ages 1 and 12 months, the test-retest correlation was .83. For ages 24 and 36 months, the correlation was .91. Across all ages, the correlation was .87 (see Bayley, 1993, p. 193-194).

Interrater Agreement

The BSID-II was administered to 51 children ranging in age from 2 to 30 months. Children were rated simultaneously by two people (the examiner, plus an additional rater who observed the assessment from nearby). The correlation between MDI scores based on the two ratings was .96 (see Bayley, 1993, p. 195).

Validity Information from Manual (Mental Scale/MDI)

Construct Validity

The construct validity of the BSID-II was addressed by examining the pattern of correlations of items with BSID-II Mental Scale and Motor Scale scores, with the expectation that each item on the BSID-II would have a higher correlation with the scale on which it was placed (the Mental Scale vs. the Motor Scale) than with the alternate scale. Bayley (1993) did not provide these correlations, but did report that "…no item consistently correlated more positively with the opposite scale. A few items were found to correlate more positively with the opposite scale at a particular age; however, at surrounding ages the items reverted to correlating more positively with the scale on which they had been placed" (pp. 29-30, 206). Further evidence provided by Bayley for the construct validity of the Mental and Motor Scales involved correlations between MDI and PDI scores. The correlation across all ages was .44. The range of correlations within age groups was .24 at 18 months to .72 at 5 months.

Concurrent Validity

Several studies comparing BSID-II MDI scores to other assessments are summarized in the manual. Across these studies, correlations range from .30 to .79 (see Bayley, 1993, p. 216-219).

- *McCarthy Scales of Children's Abilities (MSCA;* McCarthy, 1972; N=30, ages 3 years through 3 years, 6 months):
  - Verbal, r = .77.
  - Performance, r = .69.
  - Quantitative, r = .59.
  - Memory, r = .62.
  - Motor, r = .57.

  - General Cognitive Index, r = .79.
- *Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R*; Wechsler, 1989; N=40, 3 years through 3 years, 6 months):
  - Verbal IQ, r = .73.
  - Performance IQ, r = .63.
  - Full Scale IQ, r = .73.
- *Differential Ability Scales (DAS*; Elliott, 1990; N=25, 2 years, 6 months through 3 years):
  - Nonverbal Composite Score, r = .30.
  - General Conceptual Ability, r = .49.
- *Pre-School Language Scale-3 (PLS-3*; Zimmerman, Steiner, & Pond, 1992; N=66, 1 year, 6 months through 3 years, 6 months):
  - Auditory Comprehension, r = .39.
  - Expressive Communication, r = .52.
  - Total Language Score, r = .49.

Reliability/Validity Information from Other Studies

- As part of the Early Head Start analyses (see description of study below), the investigators looked at concurrent validity by calculating correlations among MDI scores and MacArthur Communicative Development Inventories (CDI) variables (Boller et al., 2002). At 14 months, the MDI demonstrated low correlations with CDI variables (ranging from .16 to .20). The correlations were moderate at 24 months (ranging from .34 to .44). In addition, the investigators found a moderate correlation of .43 between the 14-month and 24-month MDI scores. Children in both the Early Head Start program group and a control group had lower MDI scores at 24 months than at 14 months.

Comments

- Information on the reliability of the BSID-II indicate that the measure has a high degree of internal consistency, test-retest correlations are high across a short time span, and two trained raters can demonstrate high levels of agreement when assessing a child's performance within the same testing session. Taken together, these findings presented by Bayley (1993) provide strong support for the reliability of the BSID-II.
- As reported by Bayley (1993), there appears to be reasonably good evidence supporting the construct and concurrent validity of the BSID-II, although the manual did not provide sufficient information on construct validity analyses in order to allow the reader to make an independent determination of the validity of the scales. Further, BSID-II MDI scores have not always been found to correlate highly with other measures of mental development available for use with very young children. High correlations were found between BSID-II MDI scores and other measures of cognitive development, particularly the MSCA and the WPPSI-R, indicating that these measures are tapping similar constructs. Correlations were highest between the BSID-II MDI and measures of verbal and general abilities derived from the MSCA and the WPPSI-R. However, correlations of BSID-II MDI scores with DAS scores (especially Nonverbal composite scores) and PLS-3 scale scores (particularly Auditory Comprehension scores) were lower, suggesting that the constructs tapped by these two measures do not overlap as highly with mental development as assessed with the BSID-II. Similarly, results from the Early Head Start analysis provide some evidence at 24 months but little evidence at 14 months that the MDI and the CDI tap similar underlying constructs.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

The BSID and the BSID-II are among the measures used in the NICHD Study of Early Child Care (NICHD ECCRN, 1999; 2000). The study sample consists of 1,364 families in multiple cities across the United States. The study looks at the quality of child care in relation to children's outcomes across multiple domains of development. MDI scores were determined based on the original BSID at 15 months, and on the BSID-II at 24 months. The authors found that variations in child-staff ratio and group size were not related to differences in scores on the MDI (NICHD ECCRN, 1999). Caregiver education and training was also not related to MDI scores. However, a measure of positive caregiving was related to MDI scores at 24 months, although not at 15 months (NICHD ECCRN, 2000).

In another study of child care quality, Burchinal et al. (2000) studied the outcomes of 89 African American infants attending community based child care centers. The study utilized a high-risk sample—69 percent of the children were from families with incomes less than 185 percent of the poverty line, and 68 percent of the families were headed by a single parent. Children were less than 12 months old when they entered child care and were studied over the course of 4 years. The BSID was used to assess children's cognitive development at ages 1 and 2; the BSID-II was used at age 3. The authors found positive correlations between an overall rating of child care quality (the Infant/Toddler Environment Rating Scale; Harms, Cryer, & Clifford, 1990) and MDI scores at 12, 24, and 36 months, suggesting that higher quality care was related to better cognitive development. These results held even after controlling for selected child and family factors.

Intervention Study: Early Head Start is a comprehensive child development program and has a two-generation focus (Love, Kisker, Ross, Schochet, Brookes-Gunn, Paulsell, Boller, Constantine, Vogel, Fuligni, & Brady-Smith, 2002). It was started in 1995 and now serves 55,000 low-income infants and toddlers in 664 programs. There is a great deal of variation in programs, as grantees are allowed flexibility in meeting the particular needs in their communities. Program options include home-based services, center-based services, or combinations. A random assignment evaluation in 17 programs was started in 1995, and the final report was released in June 2002. The sample included 3,001 families. The BSID-II was used to assess children at 14, 24, and 36 months of age. At 24 and 36 months, children receiving services had significantly higher MDI scores than control group children. At 36 months, the program group had a mean MDI score of 91.4, while the control group had a mean score of 89.9 (however, as pointed out by the authors, Early Head Start children continued to score below the mean of the national norms.)[1]

Intervention Study: The original BSID was used in the Infant Health and Development Program (IHDP). The IHDP was an intervention project in eight cities for low birthweight premature infants and their families (Brooks-Gunn, Liaw & Klebanov, 1992). The evaluation involved 985 families and began when infants were discharged from the hospital, continuing until they were 3 years old. Infants were randomly assigned to either the intervention group or

[1] For more details, see www.acf.dhhs.gov/programs/core/ongoing_research/ehs/ehs_intro.html.

the follow-up group. Infants in both groups received medical, developmental and social assessments, as well as referrals for services such as health care. In addition, infants and families in the intervention group received three types of services: (1) home visits, (2) attendance at a child development center and (3) parent group meetings. During home visits, parents were provided with information on their children's health and development. They were also taught a series of games and activities to use to promote their children's cognitive, language and social development, and they were helped to manage self-identified problems. Beginning at 12 months of age, children attended child development centers for five days per week. The curriculum was designed to match the activities that parents were taught to carry out with their children during home visits. The last component involved parent groups, which began meeting when infants were 12 months old. Parents met every two months and were provided with information on such topics as raising children, health and safety.

The BSID was used in the evaluation at 12 and 24 months. Investigators found that children in the intervention group had higher MDI scores than children in the follow-up group (Brooks-Gunn, Liaw & Klebanov, 1992; McCormick, McCarton, Tonascia & Brooks-Gunn, 1993). In addition, the effects were the strongest for families with the greatest risk (i.e., children whose parents had a high school education or less and who were of ethnic minority status; Brooks-Gunn, Gross, Kraemer, Spiker & Shapiro, 1992).

Intervention study: The original BSID was also used in the Carolina Abecedarian Project, a child care intervention in which low-income, predominantly African American children born between 1972 and 1977 were randomly assigned to high quality center-based childcare (or a control group) from infancy until age 5. Both groups received health care and nutritional supplements. Child participants in the Abecedarian Project have been repeatedly followed-up through their school years, and a number of studies have been reported based on longitudinal data from the Project (e.g., Burchinal, Campbell, Bryant, Wasik, & Ramey, 1997; Ramey, Yeates, & Short, 1984) as well as from Project CARE, a highly similar project begin in the same community with children born between 1978 and 1980 (Burchinal et al., 1997). These studies have found that children receiving the high quality childcare intervention consistently obtained higher scores on cognitive assessments than did children in the control group, beginning with an assessment at 18 months of age using the BSID.[2]

Comments

- Findings from these studies generally suggest that BSID and BSID-II scores are affected by environmental variation, and that enriched childcare and preschool education environments may positively affect mental development (as assessed with the BSID-II) among young children living in low income families or families with other risk factors. There are studies that have not found such effects, however, and other studies have found effects that are significant but small in magnitude (e.g. Love, et al., 2002).

[2] See www.fpg.unc.edu/~ABC/embargoed/executive_summary.htm.

V. Adaptation of Measure

Bayley Short Form—Research Edition (BSF-R)

Description of Adaptation:
The BSF-R was designed to be used in the ECLS-B because the administration of the full BSID-II was deemed to be too time-consuming and complex (West & Andreassen, 2002). Both the Mental Scale and the Motor Scale of BSID-II were adapted. A set of item selection criteria was used in developing BSF-R, including psychometric properties; adequate coverage of constructs; ease of administration; objectivity of scoring; and necessity of as few stimuli as possible.

The 9-month BSF-R Mental Scale includes a core set of 13 items requiring nine administrations (in some cases, a single item can be used to code more than one response). Based on the child's performance, the interviewer may need to administer supplementary items (if the child did poorly [three or fewer correct], or if the child did well [10 or more correct]). The 18-month BSF-R Mental Scale core set includes 18 items requiring 10 administrations. The 24-month BSF-R is currently being developed.

Psychometrics of Adaptation:
The 9-month BSF-R Mental Scale correlates well with the full BSID-II Mental Scale (.74; calculated by correlating the ECLS-B field data with BSID-II publisher data). However, the correlation was not as strong for the 18-month measure (.64); work to identify the issues is ongoing.

Study Using Adaptation:
Early Childhood Longitudinal Study—Birth Cohort (ECLS-B).[3]

[3] For information on ECLS-B, see www.nces.ed.gov/ecls.

Early Childhood Measures: Cognitive

Early Childhood Measures: Cognitive

Kaufman Assessment Battery for Children (K-ABC)

I. Background Information

Author/Source

Source: Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children: Administration and scoring manual*. Circle Pines, MN: American Guidance Service. (See also Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children: Interpretive manual*. Circle Pines, MN: American Guidance Service.)

Publisher: American Guidance Service
4201 Woodland Road
Circle Pines, MN 55014
Phone: 800-328-2560
Website: www.agsnet.com

Purpose of Measure

As described by instrument publisher:
"The K-ABC is intended for psychological and clinical assessment, psychoeducational evaluation of learning disabled and other exceptional children, educational planning and placement, minority group assessment, preschool assessment, neuropsychological assessment, and research. The battery includes a blend of novel subtests and adaptations of tasks with proven clinical, neuropsychological, or other research-based validity. This English version is to be used with English-speaking, bilingual and nonverbal children" (Kaufman & Kaufman, 1983a, p. 1).

Population Measure Developed With

- The norming sample included more than 2,000 children between the ages of 2 years, 6 months and 12 years, 6 months old in 1981.
- The same norming sample was used for the entire K-ABC battery, including cognitive and achievement components.
- Sampling was done to closely resemble the most recent population reports available from the U.S. Census Bureau, including projections for the 1980 Census results.
- The sample was stratified for each 6-month age group (20 groups total) between the ages of 2 years, 6 months and 12 years, 6 months, and each age group had at least 100 children.
- The individual age groups were stratified by gender, geographic region, SES (as gauged by education level of parent), race/ethnicity (white, black, Hispanic, other), community size, and educational placement of the child.
- Educational placement of the child included those who were classified as speech-impaired, learning-disabled, mentally retarded, emotionally disturbed, other, and gifted and talented. The sample proportions for these closely approximated national norms, except for speech-impaired and learning-disabled children, who were slightly under-represented compared to the proportion within the national population.

Age Range Intended For
Ages 2 years, 6 months through 12 years, 6 months old. The subtests administered vary by the age of the child.

Key Constructs of Measure
There are two components of the K-ABC, the Mental Processing Scales and the Achievement Scale, with a total of 16 subtests. The assessment yields four Global Scales:

- *Sequential Processing Scale*: The subtests that make up this scale entail solving problems where the emphasis is on the order of stimuli.
- *Simultaneous Processing Scale*: Subtests comprising this scale require a holistic approach to integrate many stimuli to solve problems.
- *Mental Processing Composite Scale*: Combines the Sequential and Simultaneous Processing Scales, yielding an estimate of overall intellectual functioning.
- *Achievement Scale*: Assesses knowledge of facts, language concepts, and school-related skills such as reading and arithmetic.

This summary includes information on the four Global Scales. See the K-ABC summaries in the Language/Literacy and Math sections of this review compendium for information on two subtests from the Achievement Scale—the Expressive Vocabulary subtest and the Arithmetic subtest.

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
The K-ABC utilizes basals and ceilings. The child's chronological age is used to determine the starting item in each subtest. To continue, the child must pass at least one item in the first unit of items (units contain two or three items). If the child fails all items in the first unit, the examiner then starts with the first item in the subtest (unless he/she started with the first item—in that case, the subtest is stopped). In addition, there is a designated stopping point based on age. However, if the child passes all the items in the last unit intended for the child's chronological age, additional items are administered until the child misses one item.

The child responds to requests made by the examiner. The child is required to give a verbal response, point to a picture, build something, etc. Some examples of responses are:

- Repeat a series of digits in the same sequence as the examiner performed them.
- Name an object or scene pictured in a partially completed "inkblot" drawing.
- Recall the placement of pictures on a page that was exposed briefly.

- Name a well-known person, fictional character, or place in a photograph or drawing.
- Identify letters and read words.

Who Administers Measure/Training Required?

Test Administration:

- "Administration of the K-ABC requires a competent, trained examiner, well versed in psychology and individual intellectual assessment, who has studied carefully both the *K-ABC Interpretive Manual* and [the] *K-ABC Administration and Scoring Manual*. Since state requirements vary regarding the administration of intelligence tests, as do regulations within different school systems and clinics, it is not possible to indicate categorically who may or may not give the K-ABC" (Kaufman & Kaufman, 1983a, p.4).

  "In general, however, certain guidelines can be stated. Examiners who are legally and professionally deemed competent to administer existing individual tests…are qualified to give the K-ABC; those who are not permitted to administer existing intelligence scales do not ordinarily possess the skills to be K-ABC examiners. A K-ABC examiner is expected to have a good understanding of theory and research in areas such as child development, tests and measurements, cognitive psychology, educational psychology, and neuropsychology, as well as supervised experience in clinical observation of behavior and formal graduate-level training in individual intellectual assessment" (Kaufman & Kaufman, 1983a, p. 4).

Data Interpretation:

- (Same as above.)

Setting (e.g., one-on-one, group, etc.)

One-on-one.

Time Needed and Cost

Time:

- The time it takes to administer K-ABC increases with age because not all of the subtests are administered at each age. The administration time for the entire battery increases from about 35 minutes at age 2 years, 6 months to 75-85 minutes at ages 7 and above. The manuals do not provide time estimates for subtests or scales.

Cost:

- Complete kit: $433.95
- Two Manual set (*Administration and Scoring Manual* and *Interpretive Manual*): $75.95

Comments

- A Nonverbal Scale can be used with children age 4 and above who have language disorders or do not speak English. The scale consists of the Mental Processing subtests that can be administered in pantomime and responded to nonverbally.

III. Functioning of Measure

Reliability Information from the Manual

Split-Half Reliability

Because of the basal and ceiling method used in the K-ABC, split-half reliability was calculated by taking the actual test items administered to each child and dividing them into comparable halves, with odd number questions on one half and even numbers on the other. Scale scores were calculated for each half and correlated with each other, and a correction formula was applied in order to estimate reliabilities for full-length tests.

- At the preschool level (ages 2 years, 6 months to 4 years), the split-half reliability coefficients for the subtests comprising the Mental Processing Composite ranged from .72 to .89. The values at age 5 for individual subtests ranged from .78 to .92.
- At the preschool level, mean split-half reliability estimates for the subtests comprising the Achievement Composite ranged from .77 to .87. The values at age 5 ranged from .81 to .94 (see Kaufman & Kaufman, 1983b, p. 82).
- For the Global Scales, split-half reliabilities at the preschool level were as follows: Sequential Processing, .90; Simultaneous Processing, .86; Mental Processing Composite, .91; Achievement, .93; and Nonverbal, .87. At age 5, reliability estimates were as follows: Sequential Processing, .92; Simultaneous Processing, .93; Mental Processing Composite, .95; Achievement, .96; and Nonverbal, .93 (see Kaufman & Kaufman, 1983b, p. 83).

Test-Retest Reliability

The K-ABC was administered twice to 246 children, two to four weeks after the first administration. The children were divided into three age groups (ages 2 years, 6 months through 4 years; 5 years through 8 years; and 9 years through 12 years, 6 months).

- For the youngest group, test-retest correlations for the Global Scales were: Sequential Processing, .77; Simultaneous Processing, .77; Mental Processing Composite, .83; Achievement, .95; and Nonverbal, .81.
- Test-retest correlations at the level of subtests ranged from .62 to .87 for the youngest age group (see Kaufman & Kaufman, 1983b, p. 83).

Validity Information from the Manual

Construct Validity

- *Developmental changes.* In the standardization sample, raw scores on all of the K-ABC subtests, as well as the Global Scales, increase steadily with age. Kaufman and Kaufman (1983b, p. 100), describe such a pattern of age-related increases as necessary, but not sufficient, to support the construct validity of any test purporting to be a measure of achievement or intelligence.
- *Internal consistency.* The authors examined correlations between subtests and Global Scales as another assessment of construct validity. As stated by Kaufman & Kaufman (1983b, p. 100), "The homogeneity or internal consistency of a multiscore battery can be determined by correlating subtest scores with total test scores; these coefficients provide evidence of the test's construct validity."
    - At the preschool level (ages 2 years, 6 months to 4 years), correlations between the Mental Processing subtests and the Mental Processing Composite ranged from .54 to

.67. At age 5, correlations ranged from .58 to .71 (see Kaufman & Kaufman, 1983b, p. 103).
    - Correlations between Achievement Scale subtests and the Achievement Global Score ranged from .73 to .80 for the preschool-level group. At age 5, correlations ranged from .75 to .83 (see Kaufman & Kaufman, 1983b, p. 104).
- *Factor analysis*. Kaufman and Kaufman (1983b) presented the results of both principal components analyses and confirmatory factor analyses of data from the standardization sample.
    - Summarizing the results of the principal components analyses, Kaufman and Kaufman (1983b) state, "When only the Mental Processing subtests were analyzed, there was clear-cut empirical support for the existence of two and only two factors at each age level. Furthermore, orthogonal (varimax) rotation of these factors for each group produced obvious Simultaneous and Sequential dimensions" (p. 102). Specific results of analyses were presented for selected ages, including ages 3 (N=200) and 6 (N=200). At both ages, all subtests loaded most highly on the hypothesized dimension. At both ages, all factor loadings were above .40 with the exception of Face Recognition at age 3, which loaded .37 on the Simultaneous Processing dimension (p. 105).
    - When the achievement subtests were included in a second set of analyses, Kaufman and Kaufman (1983b) state that "…at ages 2 ½ and 3…only two factors should be interpreted…however, three factors produced the most sensible reduction of the data for ages 4 and above" (p. 106). At the older ages, the three factors corresponded to Sequential Processing, Simultaneous Processing, and Achievement dimensions; at the younger ages the two factors appeared to be Sequential Processing and Simultaneous Processing/Achievement. Specific results of analyses were presented for selected ages, including age 4 (N=200). At that age, all subtests loaded most highly (with factor loadings of .40 or higher) on the hypothesized dimension, with the exception of Arithmetic. Although designated an Achievement subtest, it loaded only .38 on the Achievement dimension while loading .66 on the Sequential Processing dimension (p. 105).
    - With respect to the confirmatory factor analyses, Kaufman and Kaufman (1983b) report that "The Sequential-Simultaneous dichotomy was confirmed for all age groups, and the Sequential-Simultaneous-Achievement organization of K-ABC subtests was also confirmed for all ages, including 2½- and 3-year-olds" (p. 107). However, no goodness-of-fit indices were provided.
- *Convergent and discriminant validity*. Kaufman and Kaufman (1983b) report two studies in which subtests of the K-ABC were correlated with Successive and Simultaneous factor scores on the Das-Kirby-Jarman Successive-Simultaneous Battery (Das, Kirby, & Jarman, 1975; 1979). One study involved a group of 53 learning disabled children ranging in age from 7½ to 12½; the other was a study of 38 trainable mentally retarded children ages 6 years, 3 months to 17 years, 2 months (see Kaufman & Kaufman, 1983b, p. 110).
    - The K-ABC Sequential Processing scale correlated .69 and .50 with the Das-Kirby-Jarman (D-K-J) Successive Factor in the mentally retarded and learning disabled groups, respectively, and only .27 and .32 with the D-K-J Simultaneous Factor.

  - The K-ABC Simultaneous Processing scale correlated .47 and .54 with the D-K-J Simultaneous Processing Factor and only -.11 and .12 with the D-K-J Successive Processing Factor.
  - Similarly, all K-ABC subscales correlated most highly with the predicted D-K-J Factor with the exception of Hand Movements. This subtest was expected to correlate most highly with the D-K-J Successive Factor, but correlations were nearly equal with the Success and Simultaneous factors (.45 and .42 in the trainable mentally retarded group; .30 and .31 in the learning disabled group).
- *Correlations with other tests*. A number of studies were reported by Kaufman and Kaufman (1983b) investigating associations between scores on the K-ABC and scores on other measures of cognitive functioning, achievement, or intelligence. Several of these studies using different samples were conducted to investigate the correlations between the K-ABC scales and Stanford-Binet IQ scores in preschool and kindergarten samples(see Kaufman & Kaufman, 1983b, p. 117).
  - In a kindergarten sample (N=38), correlations of K-ABC Global Scales with Stanford-Binet IQ scores ranged from .63 to .79, with the lowest correlation being Sequential Processing and the highest being the Achievement scale.
  - In a preschool sample (N=39) correlations ranged from .31 to .74, with the Nonverbal scale having the weakest correlation with Stanford Binet IQ scores and the Achievement scale having the strongest association.
  - In another preschool sample (N=28) correlations between Stanford Binet IQ scores and K-ABC scale scores ranged from .15 (with Simultaneous Processing) to .57 (with Achievement).
  - Finally, in a high-risk preschool sample (N=28) correlations ranged from .52 to .66, with the Achievement scale having the lowest association with Stanford-Binet IQ scores and the Mental Processing Composite being the most strongly associated. The high-risk sample consisted of children identified as having speech impairment, language delay, high activity level, or multiple problems involving some degree of physical disability.

Predictive Validity

Kaufman and Kaufman (1983b) reported a series of studies examining associations between K-ABC Global Scale scores and scores on achievement tests administered between 6 and 12 months later. One of these studies examined correlations between K-ABC Global scales with scores on the Woodcock-Johnson Psycho-Educational Battery (Preschool and Knowledge Clusters) administered 11 months later, in a sample of 31 normally-developing preschoolers (ages 3 years to 4 years, 11 months). The strongest correlations were between K-ABC Achievement scores and the Woodcock-Johnson Preschool and Knowledge Clusters (.73 and .84, respectively). K-ABC Mental Processing Composite scores correlated .61 and .63 with Preschool and Knowledge Cluster scores, respectively, and K-ABC Simultaneous Processing scores also correlated .61 with Knowledge Cluster scores. Other correlations between K-ABC scale scores and Woodcock-Johnson Cluster scores ranged from .33 for K-ABC Nonverbal scores correlated with Preschool Cluster scores, to .51 for Sequential Processing scores correlated with Preschool Cluster scores (see Kaufman & Kaufman, 1983b, p. 121).

Reliability/Validity Information from other Studies

- Glutting (1986) studied the K-ABC using a sample of 146 kindergartners (45 percent white, 16 percent black, and 39 percent Puerto Rican). The K-ABC Nonverbal scale was used, as many of the children were not proficient in English. Results indicated that K-ABC Nonverbal scale scores predicted classroom performance, as rated by teachers through a rating scale and assignment of grades.
- Williams, Voelker, and Ricciardi (1995) conducted a five-year follow-up study of K-ABC scores in a sample of 39 preschoolers; 10 had language impairment, 13 had behavior control deficits, 16 were developing normally. Their mean age at follow-up was 9 years, 9 months years. Children were assessed in preschool using the K-ABC full battery (Achievement scores and the Mental Processing Composite scores were analyzed separately). Follow-up measures used were the K-ABC, the Peabody Individual Achievement Test—Revised (PIAT-R), the Peabody Picture Vocabulary Test—Revised (PPVT-R), and the Test for Auditory Comprehension of Language—Revised (TACL-R). For the normally developing group, baseline K-ABC scores predicted scores on all of the outcome measures. For the behavior problems group, baseline K-ABC Achievement scores predicted PIAT-R follow-up scores. There were no significant relationships between baseline and outcome measures for the language impairment group.
- Krohn and Lamp (1989) studied the concurrent validity of K-ABC and the Stanford-Binet Intelligence Scale—Fourth Edition (SB-IV), both compared to the Stanford-Binet Intelligence Scale—Form LM (SB-LM). The sample consisted of 89 Head Start children, ranging in age from 4 years, 3 months to 6 years, 7 months, with a mean age of 4 years, 11 months. Fifty children were white and 39 were black. The authors found that K-ABC and SB-IV measures were moderately related to SB-LM measures.

Comments

- Overall, information provided by Kaufman and Kaufman (1983b) indicates that the K-ABC demonstrates strong split-half and test-retest reliabilities at the preschool and kindergarten age levels.
- Validity information provided by the authors similarly suggests that this test demonstrates good construct validity. However, information was not provided for all ages, and in some cases relevant information for making an independent evaluation of the data was not provided. Notably, results of confirmatory factor analyses relevant to understanding the goodness of fit of the 2- or 3-factor models were not provided.
- Research by Krohn and Lamp (1989) provides further support for the concurrent validity of the K-ABC.
- According to Kaufman and Kaufman (1983b), K-ABC scale standard scores, particularly Achievement scores, were predictive of preschool children's performance on the Woodcock-Johnson battery almost a year later, providing support for the predictive validity of the K-ABC. Similarly, results of studies by Glutting (1986) and by Williams, Voelker, and Ricciardi (1995) also provide support for the predictive validity of the K-ABC. Findings by William and colleagues, however, may suggest that the K-ABC is more predictive for

normally developing children. Results of that study should be viewed with some caution, however, due to the small sample size and the fact that the study did not use the Nonverbal scale, which perhaps might be a more appropriate assessment to use with language-impaired preschoolers.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

None found.

V. Adaptations of Measure

None found.

Early Childhood Measures: Cognitive

Early Childhood Measures: Cognitive

Peabody Individual Achievement Test—Revised (PIAT-R)

I. Background Information

Author/Source

Source: Markwardt, F. C. (1998). *Peabody Individual Achievement Test—Revised: Manual* (Normative Update). Circle Pines, MN: American Guidance Service.

Publisher: American Guidance Service, Inc.
4201 Woodland Road
Circle Pines, MN 55014-1796
Phone: 800-328-2560
Website: www.agsnet.com

Purpose of Measure

As described by instrument publisher:
"PIAT-R scores are useful whenever a survey of a person's scholastic attainment is needed. When more intensive assessment is required, PIAT-R results assist the examiner in selecting a diagnostic instrument appropriate to the achievement level of the subject. The PIAT-R will serve in a broad range of settings, wherever greater understanding of an individual's achievement is needed. Teachers, counselors, and psychologists, working in schools, clinics, private practices, social service agencies, and the court system will find it helpful" (Markwardt, 1998, p. 3).

According to the publisher, the uses of PIAT-R include individual evaluation, program planning, guidance and counseling, admissions and transfers, grouping students, follow-up evaluation, personnel selection and training, longitudinal studies, demographic studies, basic research studies, program evaluation studies, and validation studies.

Population Measure Developed With

- The PIAT-R standardization sample was intended to reflect students in the mainstream of education in the United States, from kindergarten through grade 12.
- A representative sample of 1,563 students in kindergarten through grade 12 from 33 communities nationwide was tested. The sample included 143 kindergartners. The initial testing was done in the spring of 1986. An additional 175 kindergarten students were tested at 13 sites in the fall of that year to provide data for the beginning of kindergarten.
- Ninety-one percent of the students were selected from public schools, and special education classes were excluded.
- The standardization was planned to have equal numbers of males and females and to have the same proportional distribution as the U.S. population on geographic region, socioeconomic status, and race/ethnicity.

Age Range Intended For

Kindergarten through high school (ages 5 through 18 years). Only the appropriate subsets are administered to any specific age group.

Key Constructs of Measure

The PIAT-R consists of six content area subtests:

- *General Information*: Measures the student's general knowledge.
- *Reading Recognition*: An oral test of reading that measures the student's ability to recognize the sounds associated with printed letters and the student's ability to read words aloud.
- *Reading Comprehension*: Measures the student's understanding of what is read.
- *Mathematics*: Measures the student's knowledge and application of mathematical concepts and facts, ranging from recognizing numbers to solving geometry and trigonometry problems.
- *Spelling*: Measures the student's ability to recognize letters from their names or sounds and ability to recognize standard spellings by choosing the correct spelling of a word spoken by the examiner.
- *Written Expression*: Assesses the student's written language skills at two levels. Level 1 is appropriate for kindergarten and first-grade students, and Level 2 is appropriate for Grades 2 through 12. Level 1 tests pre-writing skills such as copying and writing letters, words, and sentences from dictation.

Two composites scores can be calculated from the subtests. The Total Reading score is a combination of Reading Recognition and Reading Comprehension. The Total Test score is based on the first five subtests (excluding Written Expression).

Norming of Measure (Criterion or Norm Referenced)

Norm referenced.

Comments

- While some cognitive assessments address both achievement and underlying ability, the PIAT-R focuses strictly on achievement.
- The following limitations of the test are cited in the Manual:
  - The test is not designed to be used as a diagnostic test.
  - The test identifies a person's general level of achievement but is not designed to provide a highly precise assessment of achievement.
  - The items in the test reflect a cross section of curricula used across the United States and are not designed to test the curricula of a specific school system.
  - Administration and interpretation of the test scores require different skills. Users who are not qualified to interpret the scores are warned against interpreting a student's scores erroneously.

II. Administration of Measure

Who is the Respondent to the Measure?

Child. Respondents can range up to grade 12.

If Child is Respondent, What is Child Asked to Do?

Reading Comprehension, Mathematics, and Spelling subtests, and the first 16 items in the Reading Recognition subtest are multiple choice; young children are asked to point to their responses on a plate with four choices. The rest of the items use free response (either verbal or written).

Because the PIAT-R is administered to such a wide age range of respondents and contains a range of questions that vary greatly in difficulty, the examiner must determine a *critical range*. The critical range includes those items of appropriate difficulty for the student's level of achievement. Details on how to determine the critical range are provided in the PIAT-R Manual. PIAT-R utilizes basals and ceilings.

Who Administers Measure/Training Required?

Test Administration:

- Any individual who learns and practices the procedures in the PIAT-R Manual can become proficient in administering the test. Each examiner should thoroughly study Part II and Appendix A of the Manual, the test plates, the test record, and the Written Expression Response Booklet.

Data Interpretation:

- Interpretation requires an understanding of psychometrics, curriculum, and the implications of a student's performance. With these qualifications, the test can be interpreted by those with knowledge and experience in psychology and education, such as psychologists, teachers, learning specialists, counselors, and social workers, are the most likely candidates for interpreting scores

Setting (e.g., one-on-one, group, etc.)

One-on-one.

Time Needed and Cost

Time:

- There is no time limit on the test. The only timed subtest is Level II of the Written Expression subtest for which students are given 20 minutes. Typically all six subtests can be administered in one hour.

Cost:

- Complete kit: $342.95
- Manual: $99.95

Comments

- The subtests of the PIAT-R are designed to be administered in a specific order. All six subtests should be administered to ensure maximum applicability of the norms.
- If the student is young, it may be necessary to do training exercises (provided at the beginning of each subtest) to acquaint the child with pointing as the appropriate method of responding to the multiple choice questions.

III. Functioning of Measure

Reliability Information from Manual

Split-Half Reliability

For each subtest, reliability estimates were obtained by calculating correlations between the total raw score on odd items and the total raw score on even items and applying a correction formula to estimate the reliabilities of full-length tests. The manual presents results both by grade level and by age. For the kindergarten subsample, subtest reliabilities ranged from .84 to .94. The reliability for the Total Test score was .97 (see Markwardt, 1998, p.59).

Test-Retest Reliability

Students were randomly selected from the standardization sample. Fifty students were selected in each of grades kindergarten, 2, 4, 6, 8, and 10. Students were retested from two to four weeks after the initial assessment. The manual presents results both by grade level and by age. For the kindergarten subsample, test-retest correlations for the subtests ranged from .86 to .97. The coefficient for the Total Test score was .97 (see Markwardt, 1998, p.61).

Other Reliability Analyses

A total of four different reliability analyses were reported. In addition to split-half and test-retest reliabilities (summarized above), Kuder-Richardson and item response theory methods were used to estimate reliability. Results of these analyses (conducted both by grade and by age) parallel the split-half and test-retest reliability results, According to the author, these further analyses support the reliability of both the subtests and composite scores (see Markwardt, 1998, pp. 59-63).

Validity Information from Manual

Construct Validity

Intercorrelations were calculated between PIAT-R subtests and composite scores for the entire standardization sample, with separate analyses reported for selected grades (kindergarten and grades 1, 3, 5, 7, 9, and 11) and ages (5, 7, 9, 11, 13, 15, and 17 years). Correlations were found to be higher for subtests measuring similar constructs (e.g., Reading Comprehension and Reading Recognition) than for subtests tapping different constructs (e.g., Reading Comprehension and Mathematics), providing support for the construct validity of the test (see Markwardt, 1998, p. 67). Focusing on results for kindergartners (N = 143):

- Reading Recognition and Reading Comprehension correlated highly (.97) with each other and correlated .96 and .94, respectively, with Total Reading scores.
- Correlations of Reading Recognition, Reading Comprehension, and Total Reading scores with Spelling subtest scores were somewhat lower (.77, .79, and .78, respectively).
- The four language-related scales (including Spelling) had correlations that were still lower, ranging from .55 to .56, with Mathematics subtest scores.
- Correlations between the General Information subtest scores and scores on all other subtests ranged from .50 to .57.

Concurrent Validity

Scores on PIAT-R were correlated with Peabody Picture Vocabulary Test—Revised (PPVT-R; Dunn & Dunn, 1997) scores. Sample descriptions are not provided in detail, but the sample

included 44 5-year-olds and 150 6-year-olds. Correlations between PPVT-R scores and PIAT-R subtest and composite scores providing some support for the validity of the PIAT-R (see Markwardt, 1998, p. 66).

- For 5-year-olds:
  - Correlations ranged from .51 to .80, with the PIAT-R Mathematics and Spelling scales being the least correlated with PPVT-R scores, and General Information the most highly correlated.
  - The correlation between the Total Test scale and the PPVT-R was .71.
- For 6-year-olds:
  - Correlations ranged from .47 to .78, with the Spelling scale being the least associated with PPVT-R scores and the General Information scale being the most.
  - The correlation between the Total Test scale and the PPVT-R was .65.

Reliability/Validity Information from Other Studies

None found.

Comments

- The test developers present limited validity information. In particular, investigations of concurrent validity were only conducted examining relations between scores on the PIAT-R and PPVT-R, not other tests of achievement covering constructs other than language (although an appendix in the manual summarizes studies that have been conducted using the original PIAT).

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- None found that used the entire assessment. See the PIAT-R summary included with Math assessments for a summary of a child care study that used the Mathematics subscale (Blau, 1999).

V. Adaptations of Measure

None found.

Early Childhood Measures: Cognitive

Early Childhood Measures: Cognitive

Primary Test of Cognitive Skills (PTCS)

I. Background Information

Author/Source

Source: Huttenlocher, J., & Levine, S. C. (1990a). *Primary Test of Cognitive Skills: Examiner's manual*. Monterey, CA: CTB/McGraw Hill.
See also:

- Huttenlocher, J., & Levine, S. C. (1990b). *Primary Test of Cognitive Skills: Norms book*. Monterey, CA: CTB/McGraw Hill
- Huttenlocher, J., & Levine, S. C. (1990c). *Primary Test of Cognitive Skills: Technical Bulletin*. Monterey, CA: CTB/McGraw Hill.

Publisher: CTB/McGraw-Hill
20 Ryan Ranch Road
Monterey, CA 93940
Phone: 800-538-9547
Website: www.ctb.com

Purpose of Measure

As described by instrument publisher:
"The PTCS measures memory, verbal, spatial, and conceptual abilities. According to Public Law 94-142 (PL 94-142), a discrepancy between ability and achievement can be used as evidence of a learning disability. PTCS can be used with the California Achievement Tests®, Form E (CAT E) or with the Comprehensive Tests of Basic Skills, Fourth Edition (CTBS®/4) to obtain anticipated achievement information in order to screen for learning disabilities. In addition, as an ability measure, it is useful in screening for giftedness, for evidence of developmental delay, or for planning for the instructional needs of young children" (Huttenlocher & Cohen, 1990c, p. 1).

Population Measure Developed With

- Norms were derived from a random sample of kindergarten and first grade children from diverse geographic areas, socioeconomic levels, and ethnic backgrounds. There were approximately 18,000 children tested in the norming studies.
- There were two standardization periods, one in the fall of 1988 and one in the spring of 1989.
- The norming sample was stratified based on region, school size, socioeconomic status, and type of community.

Age Range Intended For

Kindergartners through first graders.

Key Constructs of Measure

The PTCS has four scales:

- *Spatial*: Abilities assessed include sequencing and spatial integration. Spatial relationships are tested in the form of sequences or patterns of shapes, and shape transformations.
- *Memory*: Abilities assessed include recall of information presented in both spatial and associative formats.
- *Concepts*: Spatial and category concepts are tested in the form of categorical and geometric relationships.
- *Verbal*: Skills assessed include object naming and syntax comprehension.

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

Comments

- The PTCS is a group-administered test designed as an initial screening device.

- The PTCS is one component of the CTB Early Childhood System. The other three components are the Early School Assessment, the Developing Skills Checklist, and Play, Learn, Grow! Instructional Activities for Kindergarten and First Grade.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
The PTCS is a paper and pencil test. Children provide responses on paper by filling in answer circles. The test is administered by one examiner who gives oral instructions for each task. It is recommended that children participate in practice exercises at least two days prior to the exam.

Different types of questions are asked for each subtest. A description of each type of question follows:

- *Spatial*:
  - *Sequences*: Children are asked to identify sequences using graphics.
  - *Spatial integration*: Children are asked to identify the picture that represents two separate objects being combined into one.
- *Memory*:
  - *Spatial memory:* Children are asked to remember where things are in a circle.
  - *Associative memory*: The examiner shows the children some pictures and tells them the names of the pictures. Then the examiner asks the children to find the picture that they named in their test booklet.
- *Concepts*
  - *Category concepts*: Children are asked to identify "things that go together, things that are the same in some way."

    - *Spatial concepts*: Children are shown a picture in a box and then asked to find the picture "that goes best with the picture in the box."
- *Verbal*:
    - *Object naming*: Children are asked to identify the things that the examiner names. For example, children are shown four pictures such as a carrot, grapes, a potato, and corn. The children are then asked to fill in the circle under the fruit.
    - *Syntax:* Children are shown similar pictures and asked to identify the picture that matches a statement read by the examiner. For example, children are asked to identify which picture shows "the boy chases the girl." The pictures show (1) a boy and a girl running side by side, (2) a boy running behind a girl, and (3) a girl running behind a boy.

Who Administers Measure/ Training Required?

Test Administration:

- The test is administered by one examiner with the assistance of proctors. It is recommended that there be one proctor for every ten children taking the test. The test should be administered to each proctor prior to administration of the actual test. The examiner and proctor(s) should also review the Examiner/Proctor Instructions.

Data Interpretation:

- Tests can be scored manually or mechanically. Hand-scoring the tests involves using the PTCS Score Key for checking. The *Norms Book* provides norms tables to convert the number of correct responses to scale scores and subsequently convert scale scores to derived scores. Alternatively, PTCS scores can be derived from a scoring service using optical scanning equipment. This service provides reports based on group and individual results. Reports can be generated specifically for performance on the PTCS or a "combination report" can be produced by combining information obtained from the PTCS along with other assessments such as the California Achievement Test (CAT E) and the Comprehensive Test of Basic Skills (CTBS/4). Scoring on the PTCS can range from raw and scale scores to norm-referenced percentile ranks, stanines, broad cognitive ability and anticipated achievement scores.

Setting (e.g., one-on-one, group, etc.)

Group. It is recommended that no more than ten kindergartners or fifteen first graders be tested at a time.

Time Needed and Cost

Time:

- Most testing sessions will be no longer than 30 minutes, although time can be varied if necessary, according to the general guidelines in the *Examiner's Manual*.

Cost:

- Administration Package: $112.50
- Hand scored forms: $76.40 for 35 forms
- Machine scored forms: $155.30 for 35 forms

- *Examiner's Manual*: $14.60
- *Norms Book*: $14.40
- *Technical Bulletin*: $16.30

Comments

- This is a paper-and-pencil test that requires respondents to be able to follow directions and fill in answers using markers. It may be necessary to administer practice tests prior to testing in order to assure that all children have attained the skills required to successfully take the test.

III. Functioning of Measure

Reliability Information from Manual

Internal Consistency

As a measure of internal consistency, split-half reliability estimates were derived using the Kuder-Richardson formula 20 (KR20) for both the kindergarten and first grade samples. For the kindergarten sample, the internal consistency coefficients were .88 for the total PTCS in the fall (N = 4127) and .87 in the spring (N = 3435). Internal reliability estimates for each of the four subtests (Spatial, Memory, Concepts, and Verbal) for fall and spring administrations of the test were somewhat lower, however, ranging from .64 to .78 (see Huttenlocher & Levine, 1990c, pp.15-16).

Test-Retest Reliability

A sample of kindergartners and first-graders were tested twice in the fall using the PTCS. Children were re-tested 2 weeks after their first assessment (sample details are not provided in the *Technical Bulletin*; the number of cases for each subtest varied—about 350 in the kindergarten sample and 370 in the first grade sample—but no explanation is provided for the variation). For the kindergarten subsample, the test-retest correlations were .65 for the Spatial subtest, .50 for the Memory subtest, .71 for the Concepts subtest, and .70 for the Verbal subtest (see Huttenlocher & Levine, 1990c, p.14).

Validity Information from Manual

Construct Validity

Intercorrelations were calculated between the PTCS subtests during the spring. The kindergarten sample at this assessment consisted of 3435 children (see Huttenlocher & Levine, 1990c, p. 9). Overall, correlations among the four subtests ranged from .35 to .54, with all subtests demonstrating high correlations with the PTCS total scores.

- Correlations between the Spatial scale and the other three scales ranged from .37 to .54; the correlation with the Memory scale was the lowest and the highest correlation was with the Concepts scale. The correlation between the Spatial scale and Total Test scores was .78.

- Correlations between the Memory scale and the other three scales ranged from .35 to .38, with the lowest being with the Verbal scale and the highest being with the Concepts scale. The correlation between the Memory scale and Total Test scores was .70.
- Correlations with the Concepts scale ranged from .38 to .54, with the Memory scale having the lowest association with the Concept scale and the Spatial scale having the strongest association. The correlation between the Concepts scale and Total Test scores was .80.
- Correlations with the Verbal scale ranged from .35 to .50, with the lowest being the Memory scale and the highest being the Concepts scale. The correlation between the Verbal scale and Total Test scores was .75.

Predictive Validity

As evidence of the predictive validity of the PTCS, the authors provide information on correlations between performance on the PTCS (as a test of abilities) and performance on two achievement tests, the CAT E (CTB/McGraw Hill, 1992) and the CTBS/4 (CTB/McGraw Hill, 1996; sample details and the timing of PTCS, CAT E, and CTBS/4 test administrations are not provided in the *Technical Bulletin*). Correlations between PTCS total test scores and CAT E scores ranged from .36 (Language Expression) to .65 (Mathematics Concepts and Applications). Correlations between the PTCS Total Test scores and with CTBS/4 section scores ranged from .50 (Word Analysis) to .64 (Reading Total). Without exception, PTCS Total Test scores correlated more highly with CAT E and CTBS/4 scores than did individual PTCS subtest scores. Further, the PTCS Memory scale consistently exhibited somewhat lower correlations across CAT E sections than did the other PTCS subscales (ranging from .19 with Language Expression to .32 with Word Analysis). This pattern was also seen for the PTCS Memory scale and the CTBS/4 sections, where correlations ranged from .23 with Word Analysis to .33 with Mathematics Concepts and Applications (see Huttenlocher & Levine, 1990c, p.10).

Reliability/Validity Information from Other Studies

None found.

Comments

- With respect to internal consistency, KR20 estimates support the reliability of total scores for the PTCS. KR20 coefficients for the four individual tests are somewhat lower than the KR20 for the total score.
- With respect to construct validity, correlations among the subtests, in conjunction with each subtest's correlations with total PTCS scores, provide some support for the authors' conceptualization that the subtests tap distinct yet related cognitive skills.
- Correlations between PTCS scores, particularly total scores, and scores on the two achievement tests provide some evidence of predictive validity. However, correlations with the Memory scale were notably lower than were other correlations with specific subtests.
- The reliability and validity information presented in the *Technical Manual* lacks specificity in terms of sample sizes and characteristics.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- None found.

V. Adaptations of Measure

- Select items from this test were adapted for use in the Early Childhood Longitudinal Study – Kindergarten Cohort (ECLS-K).[4] Data for the ECLS-K were collected for children in their kindergarten year in fall 1998, and data have been collected longitudinally. The ECLS-K assessed cognitive growth with a math and reading battery and items from the PTCS were used in measurement of these constructs. Analyses conducted at the end of this cohorts' first grade year indicated that child resources that existed at baseline (i.e., at the start of kindergarten), such as early literacy, approaches to learning, and general health, were predictive of math and reading outcomes. These relationships remained significant even after controlling for related covariates (i.e. poverty, race/ethnicity).

[4] For information on ECLS-K, see www.nces.ed.gov/ecls.

Early Childhood Measures: Cognitive

Early Childhood Measures: Cognitive

Stanford-Binet Intelligence Scale, Fourth Edition[5] (SB-IV)

I. Background Information

Author/Source

Source: Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale: Fourth Edition. Guide for administering and scoring*. Itasca, IL: The Riverside Publishing Company. (See also Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale: Fourth Edition. Technical manual*. Itasca, IL: The Riverside Publishing Company.)

Publisher: Riverside Publishing
425 Spring Lake Drive
Itasca, IL 60143-2079
Phone: 800-323-9540
Website: www.riverpub.com

Note: The fifth edition will be available in early spring 2003.

Purpose of Measure

As described by instrument publisher:

"The authors have constructed the Fourth Edition to serve the following purposes:

1. To help differentiate between students who are mentally retarded and those who have specific learning disabilities.
2. To help educators and psychologists understand why a particular student is having difficulty learning in school.
3. To help identify gifted students.
4. To study the development of cognitive skills of individuals from ages 2 to adult" (Thorndike, Hagen, & Sattler, 1986a, p. 2).

Population Measure Developed With

- One sample was used to standardize all of the subtests.
- The sampling design for the standardization sample was based on five variables, corresponding to 1980 Census data. The variables were geographic region, community size, ethnic group, age, and gender.
- Information on parental occupation and educational status was also obtained.
- The sample included 5,013 participants from ages 2 to 24. Included in this sample were 226 2-year-olds, 278 3-year-olds, 397 4-year-olds, and 460 5-year-olds.

Age Range Intended For

Ages 2 years through young adulthood.

[5] Stanford-Binet Intelligence Scale, Fifth Edition will be released in 2003.

Key Constructs of Measure
The SB-IV contains 15 subtests, covering four areas of cognitive ability:

- *Verbal Reasoning*: Vocabulary, Comprehension, Absurdities, Verbal Relations.
- *Quantitative Reasoning*: Quantitative, Number Series, Equation Building.
- *Abstract/Visual Reasoning*: Pattern Analysis, Copying, Matrices, Paper Folding and Cutting.
- *Short-term Memory*: Bead Memory, Memory for Sentences, Memory for Digits, Memory for Objects.

Subtests can be administered individually or in various combinations to yield composite Area Scores. An overall Composite Score can be calculated to represent general reasoning ability. The combination of subtests in the complete battery varies by entry level from eight to 13 subtests. Raw scores for subtests, Areas, and the Composite are converted to Standard Age Scores in order to make scores comparable across ages and across different tests.

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
SB-IV utilizes basals and ceilings within each subtest, based on sets of four items. A basal is established when the examinee passes all of the items in two consecutive sets. A ceiling is established when the examinee fails at least three out of four items in two consecutive sets.

A child is never administered all of the subtests. Guidelines for the tests to be administered are not provided based on age, but on the entry level of the examinee. Entry level is determined through a combination of the score on the Vocabulary subtest and chronological age. So, for example, children at the preschool level are typically administered the Vocabulary, Bead Memory, Quantitative, Memory for Sentences, Pattern Analysis, Comprehension, Absurdities, and Copying subtests. Other subtests will only be administered to children/adults who qualify for higher entry levels.

Examples of what the child is asked to do:

- Name a picture.
- Answer a question (e.g., "Why do people use umbrellas?").
- Say what is absurd about a picture (e.g., "square bicycle wheels").

Who Administers Measure/Training Required?
Test Administration:

- "Administering the Stanford-Binet scale requires that you be familiar with the instrument and sensitive to the needs of the examinee. Three conditions are essential to securing accurate test results: (1) following standard procedures, (2) establishing adequate rapport

between the examiner and the examinee, and (3) correctly scoring the examinee's responses" (Thorndike, Hagen, & Sattler, 1986a, p. 9).
- The manual does not provide guidelines for examiners' education and experience.

Data Interpretation:
- The manual does not specify the education and experience needed for data interpretation using the SB-IV.

Setting (e.g., one-on-one, group, etc.)
One-on-one.

Time Needed and Cost
Time:
- Time limits are not used. "Examinees vary so markedly in their test reactions that it is impossible to predict time requirements" (Thorndike, Hagen, & Sattler, 1986a, p. 22).

Cost:
- Examiner's Kit: $777.50
- *Guide for Administering and Scoring Manual*: $72.50
- *Technical Manual*: $33

Comments
- SB-IV utilizes an adaptive-testing format. Examinees are administered a range of tasks suited to their ability levels. Ability level is determined from the score on the Vocabulary Test, along with chronological age.

- At ages 4 and above, the range of item difficulty is large, so either a zero score or a perfect score on any subtest is very infrequent. However, at age 2, zero scores occur frequently on certain subtests due to an inability to perform the task or a refusal to cooperate. According to the manual, SB-IV does not discriminate adequately among the lowest 10 to 15 percent of the 2-year-old group. At age 3, SB-IV adequately discriminates among all except the lowest two percent.

III. Functioning of Measure

Reliability Information from Manual
Internal Consistency
Split-half reliabilities of the subtests were calculated using the Kuder-Richardson Formula 20 (KR-20). All items below the basal level were assumed to be passed, and all items above the ceiling level were assumed to be failed. The manual provides reliability data for every age group, but we focus on the data for ages 2 years to 5 years. At age 2, the lowest reliability found was .74, for the Copying subtest. All other split-half reliability estimates were above .80. The highest estimate was .88 for the Memory for Sentences subtest. At age 3, reliabilities ranged from .81 for Pattern Analysis and Copying, to .91 for Absurdities . At age 4, reliabilities ranged

from .81 for Vocabulary to .88 for Absurdities and Copying. At age 5, reliabilities ranged from .82 for Vocabulary to .90 for Pattern Analysis (see Thorndike, Hagen, & Sattler, 1986b, p. 39-40).

Internal consistency reliabilities were also estimated for various Area Standard Age Scores and overall Composite Standard Age Scores. These estimates were based on average correlations among subtests, and average subtest reliabilities. Reliabilities for Areas represented by multiple tests (i.e., all Areas except for Quantitative) were all above .90, with the exception of the two-test Abstract/Visual Reasoning Area at ages 2 and 3, where reliability estimates were .85 and .87, respectively. For the overall Composite, the reliabilities were consistently high; .95 at age 2, .96 at age 3, and .97 at ages 4 and 5 (see Thorndike, Hagen, & Sattler, 1986b, p. 42-44).

Test-Retest Reliability

Test-retest reliability data were obtained by retesting a total of 112 children, 57 of whom were first tested at age 5. The length of time between administrations varied from 2 to 8 months, with an average interval of 16 weeks. The age 5 subsample consisted of 29 boys and 28 girls; sixty-five percent were white, 31 percent were black, 2 percent were Hispanic, and 2 percent were Native American. For the age 5 subsample, test-retest reliabilities ranged from .69 to .78 for the subtests with the exception of Bead Memory, which had a somewhat lower test-retest correlation of .56. Test-retest correlations for the Area scores were .88 for Verbal Reasoning, .81 for Abstract/Visual Reasoning, .71 for Quantitative Reasoning, and .78 for Short-Term Memory. The reliability for the Composite score was .91 (see Thorndike, Hagen, & Sattler, 1986b, p. 46).

Validity Information from Manual

Construct Validity

A series of confirmatory factor analyses was conducted to assess whether the conceptual model underlying the SB-IV was supported (i.e., whether evidence could be found for a general ability factor as well as more specific area factors). Results of these analyses for ages 2 through 6 indicated a strong general factor reflecting performance on all tests; factor loadings for all tests were between .58 and .69 on the factor. According to the authors, there was less clear support for the four specific ability areas, with only a verbal factor (primarily influencing Vocabulary, Comprehension, and to a lesser extent Absurdities and Memory for Sentences) and a factor labeled "abstract/visual" (primarily influencing Bead memory, and to a lesser extent Quantitative, pattern Analysis, and Copying) appearing in the analysis (see Thorndike, Hagen, & Sattler, 1986b, p. 55).

Correlations were also calculated between total test, area, and composite scores. The manual presents the results for the entire sample, as well as separately by age group (see Thorndike, Hagen, & Sattler, 1986b, p. 110-113). In general, an examination of the tables provided in the technical manual for ages 2 through 5 suggests that the tests do not correlate more highly with other tests within the same area than with tests conceptually associated with different ability areas. Consistent with the factor analytic results, correlations of tests with the Composite at each age indicated that all tests correlated highly with the Composite; at age 2, correlations ranged from .56 (Bead Memory) to .72 (Comprehension, Pattern Analysis, and Quantitative); at age 3, correlations ranged from .65 (Bead Memory) to .80 (Quantitative); at age 4, correlations ranged

from .71 (Memory for Sentences) to .87 (Quantitative); and at age 5, correlations ranged from .69 (Copying) to .86 (Quantitative).

Concurrent Validity
Several studies were conducted comparing SB-IV scores to scores on other tests. We focus here on the study with the youngest sample, in which SB-IV scores were compared to Verbal, Performance, and Full Scale scores on the Wechsler Preschool and Primary Scale of Intelligence (WPPSI; Wechsler, 1967). The sample consisted of 75 participants with a mean age of 5 years, 6 months. Thirty-four children were male, 41 were female. Eighty percent of the sample was white, seven percent was black, seven percent was Asian, and the remaining six percent was classified as other race/ethnicity. With one exception, SB-IV scales were found to be highly correlated with the WPPSI scales, with the Abstract/Visual Reasoning Area demonstrating the lowest associations with the WPPSI. Correlations were as follows (see Thorndike, Hagen, & Sattler, p. 64):

- SB-IV Verbal Reasoning correlated .80 with the WPPSI Verbal Scale, and also correlated .63 with the WPPSI Performance Scale and .78 with the WPPSI Full Scale.
- SB-IV Abstract/Visual Reasoning correlated .56 with the WPPSI Performance Scale, and also correlated .46 with the WPPSI Verbal Scale and .54 with the WPPSI Full Scale.
- SB-IV Quantitative Reasoning correlated .73 with the WPPSI Full Scale, and also correlated .70 with the WPPSI Verbal Scale and .66 with the WPPSI Performance Scale.
- SB-IV Short-Term Memory correlated .71 with both the WPPSI Full Scale the WPPSI Verbal Scale, and correlated .59 with the WPPSI Performance Scale.
- The SB-IV Composite correlated .80 with the WPPSI Full Scale, .78 with the WPPSI Verbal Scale, and .71 with the WPPSI Performance Scale.

Reliability/Validity Information from Other Studies

- Johnson, Howie, Owen, Baldwin, and Luttman (1993) studied the usefulness of the SB-IV with young children. The sample consisted of 121 3-year-olds (52 girls and 69 boys). The sample included both white and black children (proportions not given), but because race contributed little to the analyses beyond socioeconomic status and HOME scores, the variable was omitted from analyses. The eight SB-IV subtests appropriate for 3-year-olds were administered, as well as the Peabody Picture Vocabulary Test—Revised (PPVT-R). The investigators found that 55 percent of the children were unable to obtain a score (that is, they did not get a single item correct) on some SB-IV subtests. Thus there may be some reason for concern when using SB-IV with very young children, although it is not clear whether the findings were specific to this particular study and circumstances of administration, or represent a general issue. In addition, SB-IV composite scores and PPVT-R scores were moderately correlated (r = .66).

- Krohn and Lamp (1989) studied the concurrent validity of the Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983) and the SB-IV, both compared to the Stanford-Binet Intelligence Scale—Form LM (SB-LM; the third edition of the assessment). The sample consisted of 89 Head Start children, ranging in age from 4 years, 3 months to 6 years, 7 months, with a mean age of 4 years, 11 months. Fifty children were white and 39 were black. The authors found that K-ABC and SB-IV scores

were moderately related to SB-LM scores, supporting the concurrent validity of SB-IV and K-ABC.

- Gridley and McIntosh (1991) explored the underlying factor structure of the SB-IV. The study utilized two samples—50 2- to 6-year-olds, and 137 7- to 11-year-olds. Altogether, 90 percent of the subjects where white, and 10 percent were black. The eight subtests appropriate for use with younger ages were administered to the younger sample. Among 2- to 6-year-olds, the authors found more support for a two-factor model (Verbal Comprehension and Nonverbal Reasoning/Visualization) or a three-factor model (Verbal Comprehension, Nonverbal Reasoning/Visualization, and Quantitative) than for the four-factor structure established by the test developers (but which, as indicated earlier, did not receive strong support in the developers' own confirmatory factor analyses).

- Saylor, Boyce, Peagler, and Callahan (2000) compared the SB-IV to the Battelle Developmental Inventory (BDI; Newborg, Stock, Wnek, 1984) with a sample of at-risk preschoolers. The sample consisted of 92 3-, 4-, and 5-year-olds who were born at very low birthweight and/or had intraventricular hemorrhage and other medical complications. The investigators found that the two measures were significantly correlated (r = .73 to .78). The authors also investigated the efficacy of the two measures in identifying children as delayed. In one set of analyses, they defined "delayed" as being one standard deviation (SD) below the mean; in a second set of analyses, they defined it as two SDs below the mean. They found that when using the more restrictive cut-off (two SDs), 100 percent of the children were identified by both measures as delayed. However, when using the less restrictive one SD threshold, the SB-IV identified only 13 percent of the children identified by the BDI as delayed. The authors suggest that the SB-IV should be used with caution when identifying children at risk for developmental delays and when determining intervention eligibility.

Comments[6]

- With respect to internal consistency reliability, data for subtest, Area, and Composite scores indicate excellent reliability of SB-IV measures—particularly for the overall Composite. The authors do caution, however, that KR-20 estimates provided for the subtests likely represent upper bounds for estimates of reliability, given that one assumption of the formula—that all items above the ceiling level are assumed to be failed—is not likely to be entirely met. The authors further recommend that overall Composite Standard Age Scores "…be used as the primary source of information for making decisions," given that these scores were found to have the highest reliability (Thorndike, Hagen, & Sattler, 1986b, p. 38).
- With respect to test-retest correlations, information provided in the Manual again provides support for the reliability of the SB-IV. As the authors note, reliabilities are higher for composited scores (i.e., higher for Area scores than for subtest scores, highest for the overall Composite).

[6] See Technical Note on criteria for reliability and validity.

- With respect to concurrent validity, the high correlation between the SB-IV Composite and WPPSI Full Scale scores provide support for the validity of the SB-IV as a measure of general cognitive ability.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- **Intervention Study:** The Stanford-Binet Form LM was used in the Infant Health and Development Program (IHDP). The IHDP was an eight-city intervention project for low birthweight premature infants and their families. The evaluation involved 985 families, beginning when infants were discharged from the hospital and continuing until they were 3 years old. Infants were randomly assigned to either the intervention group or the follow-up group. Infants in both groups received medical, developmental and social assessments, as well as referrals for services such as health care. In addition, infants and families in the intervention group received three types of services: (1) home visits, (2) attendance at a child development center, and (3) parent group meetings. During home visits, parents were provided with information on their children's health and development. They were also taught a series of games and activities to use to promote their children's cognitive, language and social development, and they were helped to manage self-identified problems. Beginning at age 12 months, children attended child development centers five days per week. The curriculum was designed to match the activities that parents were taught to carry out with their children during home visits. The last component was parent groups, which began meeting when infants were 12 months old. Parents met every two months and were provided with information on such topics as raising children, health and safety.

  The Stanford-Binet Form LM was used at 36 months. Investigators found that children in the intervention group had higher scores than children in the follow-up group (Brooks-Gunn, Liaw & Klebanov, 1992; McCormick, McCarton, Tonascia & Brooks-Gunn, 1993). Effects were strongest for families with the greatest risk (i.e., children whose parents had a high school education or less and who were of ethnic minority status; Brooks-Gunn, Gross, Kraemer, Spiker & Shapiro, 1992).

- **Intervention Study:** The Stanford-Binet Form LM was used in the Carolina Abecedarian Project at the ages of 24, 36, and 48 months (Burchinal, Campbell, Bryant, Wasik, & Ramey, 1997). The Abecedarian Project was a controlled child care intervention where children from low-income families were randomly assigned to receive high quality care or to be part of a control group from infancy until age 5.[7] Beginning at 18 months of age, children receiving the high quality care intervention consistently obtained higher scores on cognitive assessments than did the control group children, as measured by an array of cognitive measures, including the Stanford-Binet Intelligence Form L-M. Although the gap between experimental and control group scores lessened over time, differences remained significant when the children were assessed again at 12 and 15 years of age.

[7] See www.fpg.unc.edu/~ABC/embargoed/executive_summary.htm.

- **Intervention Study:** The Stanford-Binet Form LM was also used in the High/Scope Perry Preschool Study (Schweinhart, Barnes, & Weikart, 1993; Weikart, Bond, & McNeil, 1978).[8] This intervention for high risk preschool children followed participants from preschool through adulthood. Children in the experimental group who received comprehensive, high quality child care scored higher the Standford-Binet Intelligence Scale than control children who did not receive such high quality care. Effects on the Stanford-Binet intellectual outcomes dissipated as the children grew older, showing null effects at ages eight and nine, while other effects such as student retention rates remained better for the experimental group.

V. Adaptations of Measure

None found.

[8] See http://www.highscope.org/Research/PerryProject/perrymain.htm for a description of the study.

Early Childhood Measures: Cognitive

Early Childhood Measures: Cognitive

Woodcock-Johnson III (WJ III)

I. Background Information

Author/Source

Source: McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III: Technical manual*. Itasca, IL: The Riverside Publishing Company. (See also Mather, N., & Woodcock, R. W. (2001a). *Woodcock-Johnson III Tests of Achievement: Examiner's manual*. Itasca, IL: The Riverside Publishing Company; Mather, N., & Woodcock, R. W. (2001b). *Woodcock-Johnson III Tests of Cognitive Abilities: Examiner's manual*. Itasca, IL: The Riverside Publishing Company.)

Publisher: Riverside Publishing
425 Spring Lake Drive
Itasca, IL 60143-2079
Phone: 800-323-9540
Website: www.riverpub.com

Purpose of Measure

As described by instrument publisher:
The purpose of this measure is to determine an individual's cognitive strengths and weaknesses, to determine the nature of impairment, and to aid in diagnosis. The Woodcock-Johnson III (WJ III) can also be used to make decisions regarding educational programming for individual children. The authors also view it as a good research tool.

"The WJ III batteries were designed to provide the most valid methods for determining patterns of strengths and weaknesses based on actual discrepancy norms. Discrepancy norms can be derived only from co-normed data using the same subjects in the norming sample. Because all of the WJ III tests are co-normed, comparisons among and between a subject's general intellectual ability, specific cognitive abilities, oral language, and achievement scores can be made with greater accuracy and validity than would be possible by comparing scores from separately normed instruments" (McGrew & Woodcock, 2001, p. 4).

Population Measure Developed With

- The norming sample for WJ III consisted of a nationally representative sample of 8,818 subjects drawn from 100 U.S. communities. Subjects ranged in age from 2 years to 80+ years. The sample included 1,143 preschool children ages 2 years to 5 years who were not enrolled in kindergarten. An additional 304 children enrolled in kindergarten were also included in the sample.
- Participants were selected using a stratified random sampling design to create a sample that was representative of the U.S. population between the ages of 24 months and 90 years.
- Participants were selected controlling for Census region, community size, sex, race, and Hispanic origin. Other specific selection factors were included at different ages. For

preschoolers and school-age (K through twelfth grade), parents' education was controlled.
- All subjects were administered all tests from both the WJ III COG and the WJ III ACH (see description, below).

Age Range Intended For

Ages 2 years through adulthood (however, some Tests cannot be administered to younger children).

Key Constructs of Measure

The WJ III consists of two batteries—the WJ III Tests of Cognitive Abilities (WJ III COG) and the WJ III Tests of Achievement (WJ III ACH)

- The WJ III COG consists of 20 tests tapping seven cognitive factors: Comprehension-Knowledge, Long-Term Retrieval, Visual-Spatial Thinking, Auditory Processing, Fluid Reasoning, Processing Speed, and Short-Term Memory. Ten of the 20 tests are part of the Standard Battery, ten others are included in the Extended Battery which can be used for more in-depth assessment. Three of the tests from the Standard Battery, and three additional tests from the Extended Battery, are identified as supplemental. Tests can be administered individually or in various combinations to measure specific cognitive abilities.
    - In addition to individual test scores, three cognitive performance cluster scores can be constructed: Verbal Ability, Thinking Ability, and Cognitive Efficiency.
    - Using the Extended Battery, scores can also be obtained tapping each of the seven cognitive factors noted above.
    - The manual specifies three summary scores that can be obtained from the WJ III COG tests—the Brief Intellectual Ability score (based on three tests from the Standard Battery—Verbal Comprehension, Visual Matching, and Concept Formation), a General Intellectual Ability score based on seven tests in the Standard Battery, and a General Intellectual Ability score based on 14 tests in the Extended Battery.
- The WJ III ACH contains 22 tests tapping five curricular areas: Reading, Oral Language, Mathematics, Written Language, and Academic Knowledge (e.g., science, social studies). As with the WJ III COG, some tests are part of the Standard Battery, and some make up an Extended Battery. Tests can be administered individually or in combination to create cluster scores.
    - Scores that can be constructed for children ages 5 and younger include Oral Language, Broad Reading, Broad Math, Academic Skills, and Academic Applications from the Standard Battery, as well as Oral Language, Oral Expression, Listening Comprehension, Basic Reading Skills, Reading Comprehension, Math Calculation Skills, Math Reasoning, Basic Writing Skills, Written Expression, and Phoneme/Grapheme Knowledge from the Extended Battery; only Oral Language from the Standard Battery and Oral Language, Oral Expression, Listening Comprehension, and Math Reasoning are used with children beginning at age 2.
    - A Total Achievement score can be obtained for children age 5 years or older by administering nine of the Tests covering Reading, Mathematics, and Written Language.

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

Comments

- Not all of the subtests can be administered to 2-year-olds. Therefore, some composite scores can only be obtained for children 3 years of age and older, 4 years of age and older, or 5 years of age and older (depending on the composite).

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
The WJ III utilizes basals and ceilings; the rules are different for each subtest. Examples of what the respondent is asked to do:

- *For the WJ III COG:*
  - Point to the picture of a word spoken by the examiner.
  - Identify two or three pieces that form a complete target shape.
  - Listen to a series of syllables or phonemes and then blend them into a word.
  - Point to the matching shapes in a row of four or five shapes.
- *For the WJ III ACH*:
  - Identify a printed letter.
  - Listen to and recall details of a story.
  - Identify an object.
  - Solve a simple arithmetic problem.

Who Administers Measure/ Training Required?
Test Administration:

- Examiners who administer the WJ III should have a thorough understanding of the administration and scoring procedures. They should also have formal training in assessment, such as college coursework or assessment workshops.

Data Interpretation:

- Interpretation of WJ III scores requires more knowledge and experience than that required for administering and scoring the test. Examiners who interpret WJ III results should have graduate-level training in statistics and in the procedures governing test administration, scoring, and interpretation.

Setting (e.g., one-on-one, group, etc.)
One-on-one.

Time Needed and Cost

Time:

- The time needed for test administration depends on the number and combination of subtests being administered. Each subtest requires about 5 to 10 minutes. At age 2, the Standard Battery including both Cognitive and Achievement components would take between 1 and 2 hours. At age 5 the Standard Battery would take between 1½ and 3 hours.

Cost:

- Complete battery: $966.50
- Cognitive Abilities battery: $601
- Achievement battery: $444
- Manual: $52

III. Functioning of Measure

Reliability Information from Manual

Internal Consistency

Internal reliabilities were calculated in one of two ways, depending on the test. For all but the tests scored for speed and tests with multiple-point scoring systems, split-half reliability estimates were calculated by correlating total scores on odd-numbered items with total scores on even-numbered items and applying a correction formula to estimate the full-test reliabilities. Items below the subject's basal level were scored as correct while items above the ceiling level were scored as incorrect. Reliabilities for speeded tests and tests with multiple-point scored items, were calculated utilizing Rasch analysis procedures.

- *WJ III COG.*
  - Individual test reliabilities ranged from .70 (Test 12: Retrieval Fluency) to .94 (Test 17: Memory for Words) at age 2; from .76 (Test 12: Retrieval Fluency) to .98 (Test (Test 18: Rapid Picture Naming) at age 3; from .64 (Test 19: Planning) to .98 (Test 18: Rapid Picture Naming) at age 4; and from .63 (Test 19: Planning) to .98 (Test 18: Rapid Picture Naming) at age 5. Planning at ages 4 and 5 were the only two tests to have internal consistency reliability estimates below .70 (see McGrew & Woodcock, 2001, pp. 109-117).
  - Internal consistency reliabilities of the cognitive performance cluster scores (Verbal Ability, Thinking Ability, and Cognitive Efficiency) ranged from .88 to .97 across the Standard and Extended Batteries at ages 2 (Verbal Ability only) through 5.
  - For the General Intellectual Ability scale, Standard Battery, reliabilities were .96 at age 3 and .97 at ages 4 and 5 (this score cannot be obtained for 2-year-olds). For the General Intellectual Ability scale, Extended Battery, correlations were .98 at ages 4 and 5 (this score cannot be obtained for 2- or 3-year-olds). For the Brief Intellectual Ability scale, correlations were .94 at age 3, .96 at age 4, and .94 at age 5 (this score cannot be obtained for 2-year-olds; see McGrew & Woodcock, 2001, pp. 131-142).

- *WJ III ACH.*
  - Individual test reliabilities ranged from .56 to .98 at age 2 (Test 3: Story Recall r = .56; others were .82 or greater); from .60 to .97 at age 3 (Test 12: Story Recall-Delayed r = .60; others were .75 or greater); from .61 to .98 at age 4 (Test 12: Story Recall-Delayed r = .61; others were .71 or greater); and from .69 to .99 at age 5. Story Recall—Delayed had the lowest split-half reliability among the tests at all ages to which it is administered (age 3 and older). Story Recall, which had the lowest reliability at age 2, had relatively low reliabilities at other ages as well (.75, .79, and .77 at ages 3, 4, and 5, respectively). In contrast, the test with the highest reliability at every age was Letter-Word Identification (Test 1; see McGrew & Woodcock, 2001, pp. 118-129).
  - Among the cluster scores that can be derived from the WJ III ACH tests for children age 5 or younger, internal consistency estimates ranged from .81 to .97 with the exception of Written Expression; this scale, used with children ages 5 and older, had a reliability of .70 at age 5 (see McGrew & Woodcock, 2001, pp. 143-151).
  - The Total Achievement scale cannot be calculated for children under age 5. At age 5, the internal consistency reliability estimate was .93 (see McGrew & Woodcock, 2001, p. 143).

Test-retest reliability

The manual presents several different test-retest reliability analyses. One analysis that included preschool-age children examined test-retest reliabilities of nine tests from the WJ III COG and six tests from the WJ III ACH.

- A sample of 52 children, ages 2 to 7 at the time of first testing, were re-tested after less than one year; test-retest correlations ranged from .75 to .86 for WJ III COG tests and from .85 to .96 for WJ III ACH tests.
- A sample of 114 children ages 2 to 7 were retested between one and two years later, correlations ranged from .57 to .82 for WJ III COG tests and between .75 and 91 for WJ III ACH tests.
- A sample of 69 children ages 2 to 7 were retested between three and ten years later; correlations ranged from .35 to .78 for WJ III COG tests and between .59 and .90 for WJ III ACH tests (see McGrew, & Woodcock, 2001, p. 40-41).

Test-retest reliabilities were also presented for 17 WJ III ACH tests and 12 clusters in a sample of 295 to 457 individuals (depending upon the test), with the number of children ages 4 to 7 completing each test at two time points ranging from 39 to 106. Participants were re-tested one year after the initial administration. For children ages 4 to 7, test-retest reliabilities ranged from .59 (Reading Fluency) to .92 (for both Letter-Word Identification and Applied Problems). The average Total Achievement test-retest reliability from ages 4 to 7 was .96 (see McGrew, & Woodcock, 2001, p. 42-43).

Validity Information from Manual

Internal Validity

Internal structure validity was examined by investigating the extent to which WJ III tests proposed to assess similar abilities were more highly related with each other than with tests tapping different abilities. Correlations between cluster scores were examined separately for

children ages 2 to 3 and 4 to 5 within the norming sample. The expected pattern was generally found, with test clusters measuring similar constructs being more highly correlated than those measuring widely differing constructs (see McGrew & Woodcock, 2001, p. 173-174).

A series of confirmatory factor analyses were also conducted, utilizing data from the standardization sample for children ages 6 and older.

- A conceptual model underlying the development of the WJ III was compared to alternative models, including those underlying the Stanford Binet IV (Thorndike, Hagen, & Sattler, 1986a) and the Wechsler Adult Intelligence Scales—Third Edition (WAIS-III; Wechsler, 1997). According to the authors, the WJ III conceptual model "…is the most plausible explanation for the standardization data… the comparisons to alternative models indicate that simpler models of intelligence…are less plausible for describing the relationships among the abilities measured by the WJ III" (McGrew & Woodcock, 2001, p. 64). Goodness of fit indices provided in the manual indicate that the hypothesized conceptual model did demonstrate the best fit to the data of any of the models tested (see p. 198).
- In the broad factor model, with few exceptions, COG tests loaded on their expected factors, and there were relatively few cross-loadings. McGrew and Woodcock conclude that "…the cognitive tests have minimized the influence of construct irrelevant variance" (p. 64). A number of the ACH tests did cross-load onto multiple factors, however.
- Results of these analyses were also interpreted as providing support for a strong general abilities factor underlying performance on all tests; each of nine broad factors identified in the confirmatory analysis in turn demonstrated moderate to high factor loadings ranging from .55 to .93 on a higher-order general abilities factor (see McGrew & Woodcock, 2001, p.191-198).

Concurrent Validity

A study of 202 young children (mean age of 4 years, 5 months; age range from 1 year, 9 months to 6 years, 3 months) was conducted in South Carolina. Children completed all of the tests from the WJ III COG and the WJ III ACH that were appropriate for preschoolers. They were also administered the Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R; Wechsler, 1989) and the Differential Abilities Scale (DAS; Elliott, 1990). Correlations between WJ III General Intellectual Ability Scales (Extended, Standard, and Brief versions) and Cognitive Factors are presented in the manual (see McGrew & Woodcock, 2001, p. 69).

- Correlations between the WJ III General Intellectual Ability Scales and WPPSI-R Full Scale IQ scores were .74 (Extended), .73 (Standard), and .67 (Brief).
- Correlations between the WJ III General Intellectual Ability Scales and WPPSI-R Verbal and Performance IQ scores tended to be very slightly lower, ranging from .60 (for the WJ III Brief Intellectual Ability Scale correlated with WPPSI-R Verbal IQ) to .68 (for both the Extended and Standard versions of the WJ III General Intellectual Ability Scale correlated with WPPSI-R Verbal IQ scores).
- Correlations between the WJ III General Intellectual Ability Scales and DAS General Conceptual Ability scores were .73 (Extended), .67 (Standard), and .67 (Brief).
- Correlations between the WJ III General Intellectual Ability Scales and DAS Verbal Ability and Nonverbal Ability scores were somewhat lower, ranging from .53 (for the WJ III Brief Intellectual Ability Scale correlated with DAS Verbal Ability scores) to .65 (for the Extended

version of the WJ III General Intellectual Ability Scale correlated with DAS Nonverbal ability scores).

A second validity study involving 32 preschoolers (mean age of 4 years, 9 months; range from 3 years, 0 months to 5 years, 10 months) was conducted in three locations. WJ III COG tests appropriate for young children were administered, as well as the Stanford-Binet Intelligence Scale—Fourth Edition (SB-IV; see McGrew & Woodcock, 2001, p.70).

- Correlations between the SB-IV Test Composite and WJ III Scales were .71 (Extended), .76 (Standard), and .60 (Brief).
- Correlations between SB-IV Verbal Reasoning and WJ III Scales were .67 (Extended), .76 (Standard), and .68 (Brief).
- Correlations were slightly lower between SB-IV Short-Term Memory and WJ III Scales: .66 (Extended), .69 (Standard), and .55 (Brief).
- Correlations were lower still between SB-IV Abstract/Visual Reasoning and WJ III Scales: .44 (Extended), .48 (Standard), and .32 (Brief).
- The lowest correlations were between SB-IV Quantitative Reasoning and WJ III Scales: .03 (Extended), .25 (Standard), and .08 (Brief).

Reliability/Validity Information from Other Studies

- Very few studies have been published about the psychometrics of WJ III since its relatively recent publication in 2001. Many studies have been conducted on the psychometric properties of WJ-R, but we were unable to find any that are relevant to the preschool age range.

Comments

- Results presented by McGrew and Woodcock (2001) indicate that internal consistency was quite variable across the various WJ III COG tests. Although most tests had reliabilities of .70 or higher, indicating strong internal consistency, one test, Planning, had only moderate internal consistency at ages 4 and 5; it is not included in assessments for children younger than age 4. Planning is identified as a Supplemental test on the Extended Battery (i.e., it is not included in the Standard Battery). Internal consistencies of the Verbal Ability, Thinking Ability, and Cognitive Efficiency cluster scores and of the General Intellectual Ability and Brief Intellectual Ability summary scores were all high.
- Internal consistency was also variable across the WJ III ACH tests. Although internal consistency was strong (.70 or higher) for most tests at all ages, Story Recall-Delayed had only moderate internal consistency at ages 3 through 5, and Story Recall had an internal consistency below .60 at age 2. Scores involving clusters of tests, and Total Achievement summary scores all demonstrated strong internal consistency.
- Test-retest reliability information provided by McGrew and Woodcock (2001) suggest high levels of consistency in WJ III COG and WJ III ACH test scores even after a two year interval, and moderate to high levels of consistency even after more extended periods of time. It should be noted that not all of the WJ III COG tests were included in

the reported analyses, and no break-downs in the age ranges of the studies (2 years to 7 years) were available to indicate whether test-retest reliabilities are similar for the very young children within the age range, compared with the older children.

- Validity analyses investigating the internal structure of the WJ III generally indicated that the tests cohere in the expected manner and tap the proposed underlying constructs. None of these analyses included data from children under the age of 6, however, and therefore the extent to which these analyses are applicable to test performance of very young children is not known.
- Results of studies examining the concurrent validity of the WJ III indicate that children's relative performance on the WJ III is fairly consistent with their relative performance on the WPPSI-R and the DAS. Further, while SB-IV Test Composite, Verbal Reasoning, Short-Term Memory, and Abstract/Visual Reasoning scores demonstrated moderate to high correlations with the WJ III General Intellectual Ability scales, SB-IV Quantitative Reasoning scores did not. These findings generally support the validity of the WJ III as an assessment of intellectual ability. At the same time, however, it should be noted that these correlations, although high, also suggest a substantial amount of variability in children's performance across the different tests. In particular, findings reported by McGrew and Woodcock (2001) suggest that children's performance on the SB-IV and the WJ III are fairly consistent at the level of general intellectual ability (with the General Intellectual Ability scale based on the Standard Battery demonstrating the strongest associations with SB-IV scores), but that there are clearly differences in the information provided by the two tests regarding children's abilities in more specific areas.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

Studies of the quality of child care and child outcomes have generally used the WJ-R math and language subtests of the Tests of Achievement, rather than General Intellectual Ability or Total Achievement scores (see the WJ III summary included with Math measures section of this review compendium).

V. Adaptation of Measure

Spanish Version of WJ III

A Spanish version of the WJ III is available.

Cognitive (General) and Math References

Bayley, N. (1993). *Bayley Scales of Infant Development, Second Edition: Manual*. San Antonio, TX: The Psychological Corporation.

Blau, D. M. (1999). The effects of child care characteristics on child development. *Journal of Human Resources, 34*(4), 786–822.

Boehm, A.E. (1986a). Boehm Test of Basic Concepts, Revised (Boehm–R). San Antonio, TX: The Psychological Corporation.

Boehm, A.E. (1986b). Boehm Test of Basic Concepts, Preschool version (Boehm–Preschool). San Antonio, TX: The Psychological Corporation.

Boller, K., Sprachman, S., Raikes, H., Cohen, R. C., Salem, M., & van Kammen, W. (2002). *Fielding and analyzing the Bayley II Mental Scale: Lessons from Early Head Start*. Paper prepared for Selecting Measures for Young Children in Large Scale Surveys, a workshop sponsored by the Research Network on Child and Family Well-Being and the Science and Ecology of Early Development, Washington, DC.

Bracken, B. A. (1998). *Bracken Basic Concept Scale—Revised: Examiner's manual*. San Antonio, TX: The Psychological Corporation.

Brooks-Gunn, J., Gross, R.T., Kraemer, H.C., Spiker, D. & Shapiro, S. (1992). Enhancing the cognitive outcomes of low birth weight, premature infants: For whom is the intervention most effective? *Pediatrics, 89*(6), 1209-1215.

Brooks-Gunn, J., Liaw, F. & Klebanov, P.K. (1992). Effects of early intervention on cognitive function of low birth weight preterm infants. *Journal of Pediatrics, 120*, 350-359.

Brunet, O., & Lézine, I. (1951). *Le développement psychologique de la premièere enfance*. Paris: Presses Universitaires.

Burchinal, M.R., Campbell, F.A., Bryant, D.M., Wasik, B.H., & Ramey, C.T. (1997). Early intervention and mediating process in cognitive performance of children of low-income African American families. *Child Development, 68*(5), 935-954.

Burchinal, M. R., Roberts, J.E., Riggins, R., Zeisel, S.A., Neebe, E, & Bryant, D. (2000). Relating quality of center child care to early cognitive and language development longitudinally. *Child Development, 71*(2), 339-357.

Campbell, F. A., Pungello, E. P., Miller-Johnson, S., Burchinal, M., & Ramey, C. T. (2001). The development of cognitive and academic abilities: Growth curves from an early childhood educational experiment. *Developmental Psychology, 37*(2), 231-242.

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*(1), 22-24.

Carrow-Woofolk, E. (1985). *Test for Auditory Comprehension of Language- Revised Edition*. Austin, TX: Pro-Ed.

Coates, S., & Bromberg, P. M. (1973). Factorial structure of the Wechsler Preschool and Primary Scale of Intelligence between the ages of 4 and 6½. *Journal of Consulting and Clinical Psychology, 40*(3), 365-370.

CTB/McGraw Hill. (1992). *California Achievement Tests, Form E.* Monterey, CA: Author.

CTB/McGraw Hill. (1996). *Comprehensive Test of Basic Skills.* Monterey, CA: Author.

Das, J. P., Kirby, J. R. & Jarman, R. F., (1975). Simultaneous and successive syntheses: An alternative model for cognitive abilities. *Psychological Bulletin*, 82, 87-103.

Das, J. P., Kirby, J. R. & Jarman, R. F., (1979) Simultaneous and successive cognitive processes. New York: Academic Press.

Dunn, L., & Dunn, L. (1981). *Peabody Picture Vocabulary Test – Revised*. Circle Pines, MN: American Guidance Service.

Dunn, L., & Dunn, L. (1997). *Peabody Picture Vocabulary Test—Third Edition: Examiner's Manual*. Circle Pines, MI: American Guidance System.

Elliott, C.D. (1990). *Differential Ability Scales*. San Antonio, RX: The Psychological Corporation.

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.

Fenson, L., Dale, P., Reznick, S., Thal, D., Bates, E., Hartung, J., Pethick, S., & Reilly, J. (1993). *MacArthur Communicative Development Inventories: User's guide and technical manual*. San Diego, CA: Singular/Thomson Learning.

Ginsburg, H. P., & Baroody, A. J. (1990). *Test of Early Mathematics Ability, Second Edition: Examiner's manual*. Austin, TX: PRO-ED, Inc.

Glutting, J. J. (1986). Potthoff bias analyses of K-ABC MPC and Nonverbal Scale IQ's among Anglo, Black and Puerto Rican kindergarten children. *Professional School Psychology, 1*(4), 225-234.

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*(3), 237-248.

Hammill, D. D., Ammer, J. F., Cronin, M. E., Mandelbaum, L. H., & Quinby, S. S. (1987). *Quick-Score Achievement Test*. Austin, TX: Pro Ed.

Harms, T., Cryer, D., & Clifford, R. M. (1990). *Infant/Toddler Environment Rating Scale*. New York: Teachers College Press.

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*(2), 116-125.

Hresko, W.P., Reid, D.K., Hammill, D.D., Ginsburg, H.P., & Baroody, A.J. (1988). *Screening children for related early educational needs*. Austin, TX: Pro-Ed.

Huttenlocher, J., & Levine, S. C. (1990a). *Primary Test of Cognitive Skills: Examiner's manual*. Monterey, CA: CTB/McGraw Hill.

Huttenlocher, J., & Levine, S. C. (1990b). *Primary Test of Cognitive Skills: Norms book*. Monterey, CA: CTB/McGraw Hill.

Huttenlocher, J., & Levine, S. C. (1990c). *Primary Test of Cognitive Skills: Technical bulletin*. Monterey, CA: CTB/McGraw Hill.

Johnson, D. L., Howie, V. M., Owen, M., Baldwin, C. D., & Luttman, D. (1993). Assessment of three-year-olds with the Stanford-Binet Fourth Edition. *Psychological Reports, 73*(1), 51-57.

Kaufman, A. S., & Kaufman, N. L. (1983a). *Kaufman Assessment Battery for Children: Administration and scoring manual*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1983b). *Kaufman Assessment Battery for Children: Interpretive manual*. Circle Pines, MN: American Guidance Service.

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.

Laughlin, T. (1995). The school readiness composite of the Bracken Basic Concepts Scale as an intellectual screening instrument. *Journal of Psychoeducational Assessment 13*(3), 294-302.

LoBello, S. G. (1991). A short form of the Wechsler Preschool and Primary Scale of Intelligence-Revised. *Journal of School Psychology, 29*(3), 229-236.

Love, J. M., Kisker, E. E., Ross, C. M., Schochet, P. Z., Brookes-Gunn, J., Paulsell, D., Boller, K., Constantine, J., Vogel, C., Fuligni, A .S., & Brady-Smith, C. (2002). *Making a difference in lives of infants and toddlers and their families: The impacts of Early Head Start. Volume Final Technical Report.*

Markwardt, F. C. (1998). *Peabody Individual Achievement Test—Revised: Manual* (Normative update). Circle Pines, MN: American Guidance Service.

Mather, N., & Woodcock, R. W. (2001a). *Woodcock-Johnson III Tests of Achievement: Examiner's manual*. Itasca, IL: The Riverside Publishing Company.

Mather, N., & Woodcock, R. W. (2001b). *Woodcock-Johnson III Tests of Cognitive Abilities: Examiner's manual*. Itasca, IL: The Riverside Publishing Company.

McCarthy, D. (1972). *McCarthy Scales of Children's Abilities*. San Antonio, TX: The Psychological Corporation.

McCormick, M. C., McCarton, C., Tonascia, J. & Brooks-Gunn, J. (1993). Early educational intervention for very low birth weight infants: Results from the Infant Health and Development Program. *Journal of Pediatrics, 123*(4), 527-533.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III: Technical manual*. Itasca, IL: The Riverside Publishing Company.

McGroder, S. M., Zaslow, M. J., Moore, K. A., & LeMenestrel, S. M. (2000). *National evaluation of welfare-to-work strategies. Impacts on young children and their families two years after enrollment: Findings from the Child Outcomes Study*. Washington, DC: Child Trends.

Newborg, J., Stock, J.R., Wnek, L. (1984). Battelle Developmental Inventory. Itasca, IL: Riverside Publishing.

NICHD Early Child Care Research Network (1999). Child outcomes when child care center classes meet recommended standards of quality. *American Journal of Public Health, 89*(7), 1072-1077.

NICHD Early Child Care Research Network (2000). The relation of child care to cognitive and language development. *Child Development, 71*(4), 960-980.

Peisner-Feinberg, E. S., & Burchinal, M. R. (1997). Relations between preschool children's child care experiences and concurrent development: The Cost, Quality, and Outcomes Study. *Merrill-Palmer Quarterly, 43*(3), 451-477.

Pierrehumbert, B., Ramstein, T., Karmaniola, A., & Halfon, O. (1996). Child care in the preschool years: Attachment, behaviour problems and cognitive development. *European Journal of Psychology of Education, 11*(2), 201-214.

Ramey, C. T., & Campbell, F. A. (1991). Poverty, early childhood education, and academic competence: The Abecedarian experiment. In A. C. Huston (Ed.), *Children reared in poverty: Child development and public policy* (pp. 190-221). New York: Cambridge University Press.

Ramey, C.T., Yeates, K.W., & Short, E. J (1984). The plasticity of intellectual development: Insights from preventative intervention. *Child Development, 55*, 1913-1925.

Saylor, C. F., Boyce, G. C., Peagler, S. M., Callahan, S. A. (2000). Brief report: Cautions against using the Stanford-Binet-IV to classify high-risk preschoolers. *Journal of Pediatric Psychology, 25*(3), 179-183.

Schneider, B. H., & Gervais, M. D. (1991). Identifying gifted kindergarten students with brief screening measures and the WPPSI-R. *Journal of Psychoeducational Assessment, 9*(3), 201-208.

Schweinhart, L.J., Barnes, H.V. & Weikart, D.P. (1993). *Significant benefits: The High/Scope Perry Preschool Study through age 27 (Monograph of the High/Scope Educational Research Foundation, 10)*. Ypsilanti, MI: High/Scope Press.

Slosson, R. L. (1983). *Intelligence Test (SIT) and Oral Reading Test (SORT): For Children and Adults*. Los Angeles: Western Psychological.

Tellegen, A., & Briggs, P. F. (1967). Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting Psychology, 31*(5), 499-506.

Terman, L.M. & Merrill, M.A. (1973). Stanford-Binet Intelligence Scale, Form LM. Itasca, IL: The Riverside Publishing Company.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986a). *Stanford-Binet Intelligence Scale: Fourth Edition. Guide for administering and scoring*. Itasca, IL: The Riverside Publishing Company.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986b). *Stanford-Binet Intelligence Scale: Fourth Edition. Technical manual*. Itasca, IL: The Riverside Publishing Company.

Wechsler, D. (1967). *Manual for the Wechsler Preschool & Primary Scale of Intelligence.* New York: The Psychological Corporation.

Wechsler, D. (1972). *Echelle d'intelligence de Wechsler pour la période préscolaire et primaire, W.P.P.S.I.* Paris, France: Les Editions du Centre de Psychologie Appliquée.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence – Revised (WPPSI-R): Manual*. San Antonio, TX: The Psychological Corporation, Harcourt Brace Jovanovich, Inc.

Weikart, D.P., Bond, J.T., and McNeil, J.T. (1978). *Ypsilanti Perry Preschool Project: Preschool Years and Longitudinal Results through Fourth Grade*. Ypsilanti, Mich.: High/Scope Press.

West, J. & Andreassen, C. (2002, May). *Measuring early development in the Early Childhood Longitudinal Study—Birth Cohort*. Paper prepared for Selecting Measures for Young Children in Large Scale Surveys, a workshop sponsored by the Research Network on Child and Family Well-Being and the Science and Ecology of Early Development, Washington, DC.

Williams, J. M., Voelker, S., & Ricciardi, P. W. (1995). Predictive validity of the K-ABC for exceptional preschoolers. *Psychology in the Schools, 32*(3), 178-185.

Woodcock, R.W. & Johnson, M.B. (1989). *Woodcock-Johnson Psycho-Educational Battery – Revised*. Itasca, IL: Riverside Publishing.

Zimmerman, I.L., Steiner, V.G., and Pond, R.E. (1992). *Preschool Language Scale-3 (PLS-3)*. San Antonio, TX: The Psychological Corporation.

Early Childhood Measures: Language

Early Childhood Measures: Language

MacArthur Communicative Development Inventories (CDI)

I. Background Information

Author/Source

Source: Fenson, L., Dale, P., Reznick, S., Thal, D., Bates, E., Hartung, J., Pethick, S., & Reilly, J. (1993). *MacArthur Communicative Development Inventories: User's guide and technical manual.* San Diego, CA: Singular/Thomson Learning.

Publisher: Singular Publishing Group (Thomson Learning)
401 West A Street Suite 325
San Diego, CA 92101
Phone: 800-347-7707
Website: www.delmarhealthcare.com

Purpose of Measure

As described by instrument publisher:
"[The] MacArthur Communicative Development Inventories (CDIs), offer a valid and efficient means of assessing early child language…. The CDI/Words and Gestures, designed for 8- to 16-month-olds, generates scores for vocabulary comprehension, vocabulary production, and the use of gestures. The CDI/Words and Sentences, designed for 16-to 30-month olds, yields scores for vocabulary production and a number of aspects of grammatical development, including sentence complexity and mean length of child's longest utterances. The norms permit a child's scores on the major components of the inventories to be converted into percentile scores, reflecting the child's relative rank to other children of the same age and sex. The inventories are finding acceptance among practitioners and researchers in a wide array of settings" (Fenson et al., 1993, p. 2).

Population Measure Developed With

- The standardization sample included 671 families with infants and 1,142 with toddlers, who were initially contacted in response to birth announcements and pediatric mailing lists obtained in New Haven, Connecticut, Seattle, Washington, and San Diego, California. Seventy-five percent of those contacted at the first two sites returned inventories. The San Diego site had a response rate of 36 percent. Of the total of 1,813 respondents, 24 were excluded for medical reasons. There were approximately equal proportions of boys and girls in each age range.
- The demographic profile of the sample was 86.9 percent white, 4 percent black, 2.9 percent Asian/pacific islander, and 6.2 percent other.
- The education of the sample was as follows: 53.3 percent of the sample had a college degree, 24.3 percent had some college, 17.9 percent had a high school diploma, and 4.5 percent had some high school or less.

Age Range Intended For

Ages 8 months through 2 years, 6 months.

Key Constructs of Measure

Infant Form – CDI /Words and Gestures:

- *Verbal Comprehension:* Words and phrases understood by child.
- *Verbal Production*: Words the child uses.
- *Gestures:* Nonverbal communication that child uses.

Toddler Form – CDI /Words and Sentences:

- *Vocabulary Production:* Words the child uses.
- *Use of Grammatical Suffixes:* How child uses plural, possessive, progressive, and past-tense endings.

References to Past, Future and Absent Objects and People: Pace of acquisition of these displacement terms.

- *Use of Irregular Nouns and Verbs:* How often child uses these properly.

Use of Overregularized Words: How child over-extends grammatical rule (e.g., "duckses," instead of ducks).

- *Word Combinations:* Whether and what word combinations the child uses.
- *Length of Longest Sentence*: Mean length of the three longest utterances (MLU).
- *Sentence Complexity:* Forced choice of sentence examples where parents choose what sentence (of gradually increasing complexity) his or her child is most likely to use.

Norming of Measure (Criterion or Norm Referenced)

Norm referenced.

Comments

- The CDI standardization sample had a higher proportion of white children, and a lower proportion of black children than the 1990 Census, and parents were more highly educated than the national average.
- Between the ages of 8 and 12 months, children with mothers who stopped their education at high school were found to have significantly higher vocabulary comprehension scores than those children whose mothers had higher levels of education. This is a counterintuitive finding, given the generally positive relationship between parental education and language/cognitive outcomes for children. This pattern did not exist in the toddler age group.
- Although the standardization sample did vary in terms of race and education, the distribution in terms of other demographic characteristics was not described (e.g., income/poverty). The relationship between CDI scores and income was examined in subsequent work (Arriaga, Fenson, Cronan, & Pethick, 1998).
- Because infants and toddlers with Down's Syndrome or any other exceptional characteristic that could affect language development were not included, the applicability of this measure to these populations is unknown.

II. Administration of Measure

Who is the Respondent to the Measure?
Parent.

If Child is Respondent, What is Child Asked to Do?
N/A.

Who Administers Measure/Training Required?
Test Administration:
The CDI is completed by the child's parent; no training is required of an examiner.

CDI/Words and Gestures (for infants):
Words :
Parents are asked a series of questions, including very basic questions about whether the child is responding to language, and whether the child comprehends or uses particular words from a provided list (organized into 19 semantic categories and involving different parts of speech).
Gestures:
Parents are asked whether the child has ever exhibited "X" gesture in "Y" context. The contexts include: First Communicative Gestures, Games and Routines, Actions with Objects, Pretending to be a Parent, and Substitutions during Play. For instance, in the Games and Routines context, the parent might be asked if his/her child has turned to look at a toy when parent has asked, "Where's the ball?" It should be noted that all of these contexts except Substitutions During Play necessitate recognition by the parent rather than recall. Substitutions During Play requires the parent to list examples in which the child has spontaneously changed the symbolic value of an object to another during play (for instance, picking up a toy hammer and then playing it like a musical instrument).

CDI/Words and Sentences (for toddlers):
Words:
Parents are asked to fill in a vocabulary production checklist, organized into 22 semantic categories, containing various parts of speech.
The vocabulary checklist is followed by questions regarding the frequency of the child's references to the past and future, and to absent and present objects.
Parents are also asked to assess morphological and syntactic development, for instance, the use of regular plural and past tense morphemes, and whether the child has begun to use common irregular plural nouns and irregular verbs. This is measured through parent response to a list of overregularized plural nouns (e.g., "teethes," "blockses") and verbs (e.g., "blowed," "sitted") to identify whether the child uses these forms.
Sentences:
The sentences section focuses on multiword utterances, and parents are asked to choose which of 37 increasingly complex sentences best reflects use by his/her child.
Parents are asked to write down three of the child's longest sentences.

Data Interpretation:
Results are to be interpreted by a researcher familiar with the CDI, not by the parent.

Setting (e.g., one-on-one, group, etc.)
One-on-one.

Time Needed and Cost
Time:
20 to 40 minutes, depending on the child's communicative skill.

Cost:
$212.95

Comments
- Relying on parental report has advantages and disadvantages. Parents have access to more information about the child's everyday language than might be elicited in an assessment situation. Parental report also eliminates the necessity of exposing very young children to testing situations. Yet there are issues with the reliability of parental report (as noted below).
- Particularly with infants, it is sometimes difficult for parents to distinguish words that infants use from the words that they truly have understanding of. The CDI does not require that parents distinguish between use and comprehension for infants.

III. Functioning of Measure

Reliability Information from Manual
Internal Consistency
- Coefficient alphas were examined to establish internal consistency for Vocabulary Comprehension, Vocabulary Production, and Gestures scales derived from the infant form, and for Vocabulary Production and Sentence Complexity composites derived from the toddler form. Both the infant and toddler form vocabulary scales showed the highest reliabilities with alphas of. 95, .96, and .96 for infant form comprehension, infant form production and toddler form production, respectively (see Fenson, et al., 1993, p. 67). Internal consistency is not provided for the 6 remaining toddler scales.

The infant form Gesture scale is comprised of categories in which the gesture might occur (i.e., context of the gesture) and when all of these categories were collapsed they showed lower internal consistency, with an alpha coefficient of .39. Scores for three of these categories (First Communicative Gestures, Actions with Objects, and Imitating Adults) were found to be highly correlated. A scale comprised of these three categories had an alpha coefficient of .79. Scores for the two remaining categories (Games and Routines, and Pretending to be a Parent), were highly correlated (.69), but were not correlated with the other three (see Fenson, et al., 1993, p. 67).

- Internal consistency for the toddler form Words and Sentence Complexity scale was .95 between the three subscales (i.e. bound morphemes, functor words, and complex sentences).

Test-Retest Reliability
Parents of 500 children re-evaluated them 6 (+/- 2) weeks after completion of the CDI. Correlations were computed between scores for the two testing periods for children at different months of age. Correlations ranged from .8 to .9 in both the Vocabulary Production and Comprehension sections; correlations for Gestures ranged from about .6 to .8 (see Fenson, et al., 1993, p. 68). Test-retest information is not reported for any of the toddler scales.

Validity Information from Manual

Content Validity

- The items for each subscale were derived from developmental literature (see Fenson, et al., 1993 for citations) and comments provided by parents.

Convergent Validity

- Patterns of growth found on the CDI were found to be similar to reports in the research literature.
- With few exceptions (such as a somewhat elevated parental report of receptive vocabularies of children at 8 months), the variation at each age group made intuitive sense given what is currently known about communicative development.
- As asserted by the author, parental observation of grammar mapped onto research reports regarding development. For instance, parents correctly observed relative time of onset for word combinations, as well as acceleration in grammatical complexity between the ages of 8 and 30 months, the sequence of emergence of specific grammatical morphemes (e.g., noun inflections come before verb inflections, and irregular past-tense verbs generally come before regular past tense verbs), and the age of onset and relative infrequency of overgeneralizations (e.g., incorrect forms like "foots" or "eated").

Concurrent Validity

- Correlations between the CDI selected scales and other measures of vocabulary were strong. Correlations between the CDI Words and Sentences Inventory and the Expressive One-Word Picture Vocabulary Test (EOWPVT; Brownell, 2000) ranged from .73 to .85, with the strongest relationship found for a sample of older, language impaired children. An earlier version of the CDI vocabulary checklist (reported to be comparable to the one currently used) was shown to be correlated with the Bayley Language Subscale (.33 -.63) in various samples of children. The weakest correlation was found within a sample of preterm infants and strongest among a full term sample. No other CDI scales were compared to standardized measures.
- To assess the validity of the Gestural scale of the CDI, a group of low gesture (N = 16) and high gesture (N = 18) children, as reported in the norming sample, were assessed using laboratory methods (i.e., spontaneous symbolic play, elicitation of recognitory gestures, forced word choice comprehension tasks, and the Snyder Gestural Communication Task) 3 to 5 months after their parents finished their initial CDIs (Thal & Bates, 1988). Those who were designated as high or low on the CDI Gestrual scale were also found within the same designation on the symbolic play and recognitory gesture laboratory measures. Similar results were shown for vocabulary comprehension scores (see Fenson, et al., 1993, p. 71-74).

In various samples, correlations of the CDI measure of Sentence Complexity and a laboratory observation of Mean Length of Utterance (Miller, 1981) ranged from .62 to .88. Though the correlation coefficients show a range, and sample sizes are similar for the three studies cited, the manual reports a significant p value only for the lowest r, .62. This relationship is strong and was found in a sample of older, language impaired children. It is assumed that because of the comparable size of the r values and the fact that alpha was set at .01, that some, if not all, of the other correlations might have been significant at the .05 level.

- In the same samples mentioned above, correlations of the CDI measure of Three Longest Sentences with a laboratory measure of Mean Length of Utterance ranged from .60 to .77. Again, a p value is only reported for the smallest correlation coefficient, and it is unknown whether the other two correlations are significant (see Fenson, et al., 1993, p. 75).

Predictive Validity
A subsample of the norming sample was given the CDI to finish 6 months after they first completed it. Correlations between time 1 and time 2 scores were examined for subgroups of children in small age ranges (e.g., 17 to 19 months; see Fenson, et al., 1993, p. 75).

For the 288 children who stayed within the toddler age range (and thus whose parents completed the same CDI form twice):
Correlations between Time 1 and Time 2 vocabulary scores were .71. In order to determine whether child age affected the stability of children's vocabulary scores from Time 1 to Time 2, the sample was broken down into 1-month age groups and the across-time correlations for the groups were compared. These correlations did show some variability based on child age, but were high throughout.
There was a significant correlation overall between grammatical complexity Time 1 and Time 2 scores, .62. When correlations within 1-month age groupings were examined, the correlation was not significant at 16 months (-.16). At 17 months and beyond, all correlations were significant; from 17 to 19 months correlations ranged from .47 to .50; correlations between 20 and 24 months were even stronger, ranging from .60 to .65.

For the 217 children who moved from infancy to toddlerhood (and thus whose parents completed different CDI forms at the two time points):
Significant correlations were found between scores from the Words and Gestures Inventory and the Words and Sentences Inventory. Correlations ranged from .38 to .73 with a median of .69 across ages.

For the 62 children who moved from younger to older infancy (and thus whose parents completed the same CDI form twice):
Correlations between the two time points were .44 (vocabulary comprehension), .38 (vocabulary production), and .44 (total gestures; see Fenson, et al., 1993, p. 75-77).

Reliability/Validity Information from Other Studies
Feldman, Dollaghan, Campbell, Kurs-Lasky, Janosky, and Paradise (2000) raised methodological concerns, including concern about the relative homogeneity of the norming

sample, appropriateness for lower income and racially diverse samples, and extent of stability of scores, especially at younger age ranges. This study found significant mean differences in CDI-Words and Sentences scores for race, maternal education, and health insurance status (proxy for income).

Fenson and colleagues responded in detail to the issues raised (see Fenson et al., 2000 for full discussion). They note, for example, that the findings of differences by race, maternal education and health insurance status may be substantive findings rather than reflecting problems with the measure, and that stability over time in the youngest children has not been found to be higher with other measurement approaches.

Comments

- Further work would help to clarify whether the Infant Gesture domain is best captured by a single scale or by two separate scales.

With regard to test-retest reliability, correlations were slightly lower for Gestures than for Vocabulary Production, but still fall in a high range for reliability.
An examination of the across-time correlations reported by Fenson, et al. (2000) suggest that predictive validity increased with increasing age across infancy and toddlerhood. Although significant, cross-age correlations were lowest within the infancy period, were somewhat higher among children who transitioned from infancy to toddlerhood, and were highest overall among the oldest toddlers.
Though reliability and validity information is available for some of the scales (most often the infant and toddler Vocabulary scales, the infant Gesture scale and the toddler Sentence Complexity scale) similar information is not reported for many of the other scales within the CDI.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

The NICHD Study of Early Child Care (2000) found several associations between child care-related variables and the various scales of the CDI. For example, an observational measure of language stimulation in the caregiving environment at both 15 and 24 months predicted CDI vocabulary production and sentence complexity scores at 24 months.

V. Adaptations of Measure

Adaptations are available in American Sign Language, Austrian-German, Basque, Chinese (Mandarin and Cantonese), Croatian, British English, Finnish, French (Canadian), Hebrew, Icelandic, Italian, and Spanish (Cuban and Mexican).

Early Childhood Measures: Language

Early Childhood Measures: Language

Expressive One-Word Picture Vocabulary Test (EOWPVT)

I. Background Information

Author/Source

Source: Brownell, R. (2000). *Expressive One-Word Picture Vocabulary Test: Manual*. Novato, CA: Academic Therapy Publications.

Publisher: Academic Therapy Publications
20 Commercial Boulevard
Novato, CA 94949
Phone: 800-422-7249
Website: www.academictherapy.com

Purpose of Measure

As described by instrument publisher:
"The EOWPVT provides a measure that reflects the extent of an individual's vocabulary that can be accessed and retrieved from memory and used to produce meaningful speech. It is a measure that depends on a number of component skills and has implications regarding an individual's cognitive, language, and academic progress. The EOWPVT has a number of specific uses: Assessing the extent of spoken vocabulary (compared to norm for age), Assessing cognitive ability (only peripherally), Diagnosing reading difficulties, Diagnosing Expressive Aphasia (because this was normed with its sister assessment of receptive language, the significance and frequency of differences between the two can be used toward this purpose and directions for this comparison are provided), Preschool and Kindergarten screening tool, Evaluating an English learner's vocabulary, Monitoring growth, Evaluating program effectiveness" (Brownell, 2000, pp. 14-15).

Population Measure Developed With

- 2,327 children were included in the norming sample for this measure and ranged in age from 2 years through 18 years, 11 months.
- Characteristics of the sample in terms of region, race/ethnicity (Asian, black, Hispanic, white, other), gender, parental education level, urban versus rural residence, and disability status (no disability, learning disability, speech/language disorder, mental retardation, other) closely matched that of the U.S. Census figures available in 1998.
- Norming sample participants were only included if their primary language was English.

Age Range Intended For

Ages 2 years through 18 years, 11 months.

Key Constructs of Measure

- The EOWPVT measures expressive vocabulary, and "requires the individual to name objects, actions, and concepts that range from familiar to obscure and in this way

provides an assessment of how the individual's expressive vocabulary compares to what is expected of the individual at a particular age level" (Brownell, 2000, p. 14).
- Because the EWOPVT was created to be used with its sister test of receptive vocabulary, the Receptive One-Word Picture Vocabulary Test (ROWPVT), significant discrepancies between the ratings on these two tests may be used for measuring Expressive Aphasia.

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

Comments
- Concern has been raised about the appropriateness of using the EWOPVT with Hispanic-American populations, despite the diversity of the sample with which the measure was developed, and despite the availability of a Spanish version of the assessment (Pena, Quinn, & Iglesias, 1992).

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
Once a basal is established, the child is presented with a series of pictures of objects, actions and concepts. A prompt is given for each picture. For example, the prompt for a picture of pyramids might be, " What are these?" The prompt for an illustration including cherries, an apple and a pear might be, "What word names all of these?" Cues may be given if the child is not attending to the feature of the picture that is intended. For example, in response to a picture of a foot with an arrow pointing to the toe, if a child says "foot," the examiner can point to the part of the picture the arrow is indicating (the toe), and ask, "What's this?"

Who Administers Measure/Training Required?
Test Administration:
- The EWOPVT is usually administered by someone with a relevant background (e.g., speech pathologist, psychologist, learning specialist). However, with training and supervision, it can be administered by someone without such a background.

Data Interpretation:
- Interpretation of scores requires familiarity with and appropriate use of statistical terms such as confidence intervals, standard scores, and percentile ranks.

Setting (e.g., one-on-one, group, etc.)
One-on-one.

Time Needed and Cost

Time:

- The test is not timed, but it usually takes 10 to 15 minutes to administer and 5 minutes to score.

Cost:

- Complete kit: $140

Comments

- The EOWPVT is a relatively inexpensive and brief measure to administer.

III. Functioning of Measure

The standardization sample consisted of 2,327 individuals from a larger group of 3,661 individuals who were administered the test in the standardization study. Reliability and validity were examined with data from the standardization study.

Reliability Information from Manual

Internal Consistency

Coefficient alphas ranged from .93 to .98, with a median of .96 across different age groups (age ranges of 1 year were examined between ages 2 and 14; age groups of 15-16 and 17-18 were also examined; see Brownell, 2000, p. 63).

Split-Half Reliability

Split-half coefficients ranged from .96 to .99, with a median of .98 (see Brownell, 2000, p. 63).

Test-Retest Reliability

226 examinees were retested an average of 20 days after first testing by the same examiner. Corrected test-retest correlations ranged from .87 to .97 for different age groupings, with a coefficient of .90 for the full sample (see Brownell, 2000, p. 65). Reliability increased with age, but correlations were strong even for the youngest children (2 years through 4 years).

Interrater Reliability

Thirty scoring sheets were randomly selected from the standardization sample, two from each of the 15 age levels. On the scoring sheets, it was possible to see items marked right or wrong, but not basals, ceilings or raw scores. Four scorers (two of whom were experienced in scoring the test and two of whom were not) calculated raw scores for each scoring sheet. Their scores were compared to computer scoring of the sheets. Agreement across all scorers was 100 percent (see Brownell, 2000, p. 64-65).

The reliability of response evaluation (i.e., the consistency in scoring an individual's response as right or wrong) was also examined. Using the same set of 30 sheets, the original examiners were asked to write the examinee's actual word response next to the item number on the sheet. All markings indicating if an item was scored right or wrong were removed, and a trained examiner reviewed and re-scored all items based on the responses that were recorded on the score sheet by the original examiner. A total of 2,508 responses were examined in this way. There was 99.4

percent agreement on scoring of the items in relation to the original scoring (see Brownell, 2000, p. 65).
In a test of the reliability of administration, 20 children ranging in age from 3 years to 17 years, 6 months were each tested by two different examiners. Following the administration, the protocols were scored by a single examiner. The corrected correlation between scores from the two protocols was .93 (see Brownell, 2000, p. 65-66).

Validity Information from Manual

Concurrent Validity

Corrected correlations between the EOWPVT and other tests of vocabulary (Expressive Vocabulary Test (Williams, 1997); PPVT-R (Dunn & Dunn, 1981) and PPVT-III (Dunn & Dunn, 1997); Receptive One-Word Vocabulary Test (Brownell, 2000); Test of Language Development (Newcomer & Hammill, 1997); WISC-III Vocabulary (Weschler, 1991); Stanford-Binet Intelligence Scale—Fourth Edition (Thorndike, Hagen, & Sattler, 1986); California Achievement Test—Fifth Edition (CTB/McGraw-Hill, 1992); Metropolitan Achievement Test—Seventh Edition (Harcourt Brace Educational Measurement, 1992); and Stanford Achievement Test—Ninth Edition (Harcourt Brace Educational Measurement, 1996)) ranged from .67 to .90 with a median of .79 (see Brownell, 2000, p. 71).

Construct Validity

There was a correlation of .84 between age and raw score for expressive vocabulary, a finding in keeping with the assumption that older individuals have larger expressive vocabularies (see Brownell, 2000, p. 73).

Correlation of the EOWPVT with the Otis-Lennon School Ability Test—Seventh Edition (OLSAT; Otis & Lennon, 1995), a measure of abstract thinking and reasoning from which both verbal and nonverbal scores are derived, was examined in a sample of 40 children and adolescents, ranging in age from 7 to 18. The corrected correlation between the EOWPVT and the OLSAT verbal score (.88) was higher than the correlation between the EOWPVT and the OLSAT nonverbal score (.71; see Brownell, 2000, p. 73).

Corrected correlations were examined between the EOWPVT and five broad measures of language focusing on "language in connected discourse" [the Clinical Evaluation of Language Functions (CELF-3; Semel, Wiig & Secord, 1995); the Oral and Written Language Scales (OWLS; Carrow-Woolfolk, 1995); the Preschool Language Scales (PLS-3; Zimmerman, Steiner & Pond, 1992); the Test for Auditory Comprehension of Language, Revised (TACL-R; Carrow-Woolfolk, 1985); and the Test of Language Development (TOLD-P:3; Newcomer & Hammill, 1997)]. Children included in this analysis ranged in age from 2 to almost 10, depending upon the measure being considered. The sample used in the analysis of the PLS-3 was the only one to include 2-year old children.

Corrected correlations between the EOWPVT and three of the five criterion measures where total language scores were reported ranged from .71 to .85, with a median of .76 (see Brownell, 2000, p. 74). The subtests of all five of the criterion measures were compared to EOWPVT scores and showed very slight variation in the relationships between the subtest scores of the

criterion measures and the EWOPVT. Correlations ranged from .64 with the PLS-3 Expressive Language subtest to .87 with both the Expressive Language and Oral Expression subtests of the CELF-3 and OWLS, respectively.

EWOPVT scores were significantly lower than average for individuals who had mental retardation, autism, language delay, expressive/receptive language disorders, behavioral disorders, learning disabilities, hearing loss, and auditory processing deficits (see Brownell, 2000, p. 77).

Scores on the EOWPVT were correlated with reading and language scores from the following achievement tests: California Achievement Test—Fifth Edition; Metropolitan Achievement Test—Seventh Edition; Stanford Achievement Test—Ninth Edition; and Woodcock-Johnson—Ninth Edition. Corrected correlations ranged from .58 to .86 (see Brownell, 2000, p. 76).

The correlation between scores on the Expressive and Receptive One Word Vocabulary Test (uncorrected) was .75 (see Brownell, 2000, p. 75).

Comments

- Regarding concurrent validity, correlations between the EOWPVT and other measures of expressive language were similar to correlations between the EOWPVT and measures of receptive language (medians: expressive, .81; receptive, .76), providing some supporting evidence for the validity of EOWPVT as a measure of language development, but less evidence of the distinctiveness of the constructs of receptive and expressive language as assessed with the EOWPVT and other measures. The authors contend that some of this unique variance might be due to the varying formats of the test.
- Among the three criterion measures that included expressive language subtests, the expressive language subtests were the most highly correlated with EOWPVT scores for two (i.e., CELF-3 and the OWLS). It is noted that these differences were not tested for significance, and in the remaining criterion with an expressive subtest, this relationship was found to be the weakest (i.e., PLS-3). This finding helps to substantiate the construct validity of this measure.
- In regard to the relationship between EOWPVT scores and the scores on various achievement tests, it is noted that of the three tests that included separate reading and language scales, correlations with the EWOPVT did not vary greatly between scales. The largest difference between correlation coefficients in these scales was in the opposite direction expected (SAT-9, Reading, .71; and Language, .58).

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- **Intervention Study**: Pena, Quinn, and Iglesias (1992) explored the possibility of cultural bias in tests of language that require the child to label pictures. They found that, in a sample of Hispanic Head Start participants, the EOWPVT did not differentiate those with true language delay from those who came from families that used non-specific labels for activities and objects. The authors provided an intervention for both the language

delayed and the non-language-delayed children who did badly on the EWOPVT, consisting of multiple activities that stressed the act of labeling objects and behaviors. They then gave a post-test with the EOWPVT. While both groups benefited from the intervention, the non-language delayed Head Start students experienced larger gains than the language delayed students.

- In a study of how the type of child care (e.g., licensed center, licensed family child care, unlicensed family child care), the quality of care at home and in child care, and familial traits predicted language outcomes (EWOPVT), Goelman and Pence (1987) found that higher quality in licensed center and licensed family care predicted better language outcomes, even after background characteristics had been taken into account.

Comments

- The test requires explicit labeling, something that may not be emphasized in all cultural contexts.

V. Adaptations of Measure

Spanish-Bilingual Version

Description of Adaptation

Source: Brownell, R. (2000). *Expressive One-Word Picture Vocabulary Test–Spanish-Bilingual Edition*. Novato, CA: Academic Therapy Publications.

"This edition offers an assessment of expressive vocabularies of individuals who are bilingual in Spanish and English. By permitting examinees to respond in both languages, this test assesses total acquired vocabulary. The test is co-normed on a national sample of Spanish-bilingual individuals ages 4-0 through 12-11. Record forms include acceptable responses and stimulus words in both languages. The manual includes administration instructions and national norms" (www.academictherapy.com, 5/28/02).

Early Childhood Measures: Language

Early Childhood Measures: Language

Kaufman Assessment Battery for Children (K-ABC), Expressive Vocabulary Subtest

I. Background Information

Author/Source

Source: Kaufman, A. S., & Kaufman, N. L. (1983a). *Kaufman Assessment Battery for Children: Administration and scoring manual*. Circle Pines, MN: American Guidance Service. (See also Kaufman, A. S., & Kaufman, N. L. (1983b). *Kaufman Assessment Battery for Children: Interpretive manual*. Circle Pines, MN: American Guidance Service.)

Publisher: American Guidance Service
4201 Woodland Road
Circle Pines, MN 55014
Phone: 800-328-2560
Website: www.agsnet.com

Purpose of Measure

A summary of K-ABC is also included within the Cognitive Assessment section of this compendium. Some of the information in the two summaries is the same, but in this summary we provide more detailed information on the subtest related to language.

As described by instrument publisher:
"The K-ABC is intended for psychological and clinical assessment, psychoeducational evaluation of learning disabled and other exceptional children, educational planning and placement, minority group assessment, preschool assessment, neuropsychological assessment, and research. The battery includes a blend of novel subtests and adaptations of tasks with proven clinical, neuropsychological, or other research-based validity. This English version is to be used with English-speaking, bilingual and nonverbal children" (Kaufman & Kaufman, 1983a, p. 1).

Population Measure Developed With

- The norming sample included more than 2,000 children between the ages of 2 years, 6 months and 12 years, 6 months old in 1981.
- The same norming sample was used for the entire K-ABC battery, including cognitive and achievement components.
- Sampling was done to closely resemble the most recent population reports available from the U.S. Census Bureau, including projections for the 1980 Census results.
- The sample was stratified for each 6-month age group (20 groups total) between the ages of 2 years, 6 months and 12 years, 6 months, and each age group had at least 100 children.
- The individual age groups were stratified by gender, geographic region, SES (as gauged by education level of parent), race/ethnicity (white, black, Hispanic, other), community size, and educational placement of the child.

- Educational placement of the child included those who were classified as speech-impaired, learning-disabled, mentally retarded, emotionally disturbed, other, and gifted and talented. The sample proportions for these closely approximated national norms, except for speech-impaired and learning-disabled children, who were slightly under-represented compared to the proportion within the national population.

Age Range Intended For
Ages 2 years, 6 months through 4 years, 11 months. It should be noted that the age range for the Expressive Vocabulary subtest is different than the age range for the K-ABC in its entirety, which extends into early adolescence.

Key Constructs of Measure
The K-ABC Expressive Vocabulary subtest measures the child's ability to state the correct names for objects pictured in photographs, demanding recall ability and verbal expression rather than recognition and receptive skills. "From the perspective of language development, recognition ability is acquired prior to recall ability, and the latter skill is more abstract and complex than the former" (Kaufman & Kaufman, 1983a, p. 51).

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

Comments
- In addition to the norming sample described above, the K-ABC had a supplementary "Sociocultural Norming Program" that included the addition of 496 black children and 119 white children, to increase the total of each group to 807 and 1,569, respectively.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
K-ABC utilizes basals and ceilings. The child's chronological age is used to determine the starting item in each subtest. To continue, the child must pass at least one item in the first unit of items (units contain two or three items). If the child fails all items in the first unit, the examiner then starts with the first item in the subtest (unless he/she started with the first item—in that case, the subtest is stopped). In addition, there is a designated stopping point based on age. However, if the child passes all the items in the last unit intended for the child's chronological age, additional items are administered until the child misses one item.

- For the Expressive Vocabulary subtest, the examiner presents the child with photograph stimuli, and the child is asked to give the names of objects verbally.

Who Administers Measure/Training Required?

Test Administration:

- The administration and interpretation of the K-ABC requires a competent, trained examiner, well-versed in individual intellectual assessment, including, most notably, the Stanford-Binet Intelligence Scale. Examiners are also expected to have a good background in the theory and practice of child development, test and measurement, cognitive psychology, educational psychology, neuropsychological development, as well as supervised experience in clinical observation and graduate-level training in individual intellectual assessment.

Data Interpretation:

- (Same as above.)

Setting (e.g., one-on-one, group, etc.)

One-on-one.

Time Needed and Cost

Time:

- The administration of this subtest is not timed. Though the time needed for individual subtests is not explicitly given in the Manual, based on the average time that it takes a child between the ages of 2 years, 6 months and 5 to take the entire battery (approximately 45 minutes), the time needed for the Expressive Vocabulary subtest is probably less than 10 minutes.

Cost:

- Complete kit: $433.95 (individual subtest cannot be purchased separately)
- Two Manual set (*Administration and Scoring Manual* and *Interpretive Manual*): $75.95

Comments

- The training that the manual indicates for administering and interpreting the K-ABC applies to the entire battery rather than to this particular subtest.
- The K-ABC Achievement Scale is generally thought of as a whole and is administered as such. The Expressive Vocabulary subtest is generally not administered separately.

III. Functioning of Measure

Reliability Information from Manual

Split-half Reliability

Split-half reliability was examined for the Expressive Vocabulary subtest (using odd and even items). A Rasch-Wright model (i.e., IRT) was used to correct for assumptions inherent with scoring below the basal and above the ceiling, and reliability coefficients for three age groups (2 years, 6 months through 2 years, 11 months; 3 years through 3 years, 11 months; and 4 years through 4 years, 11 months) ranged from .80 to .89, with a median of .85 (see Kaufman & Kaufman, 1983b, p. 82).

Test-Retest Reliability
The K-ABC was administered twice to 246 children, two to four weeks after the first administration. The children were divided into three age groups (ages 2 years, 6 months through 4 years; 5 years through 8 years; and 9 years through 12 years, 6 months), with only the youngest age group receiving the Expressive Vocabulary subtest. A total of 72 children took this subtest twice. The corrected test-retest correlation was strong, .86 (Kaufman & Kaufman, 1983b, p. 85).

Validity Information from Manual
Construct Validity
All of the K-ABC subtests, as well as composite scores, have shown a "clear-cut and consistent relationship to chronological development" (Kaufman & Kaufman, 1983, p. 100). Correlations between the Expressive Vocabulary subtest and the Achievement Global Scale ranged from .73 to .82, varying by age (2 years, 6 months through 4 years, 11 months). All correlations were within the high range, but correlations for those within 3 years to 3 years, 11 months age were slightly lower than both the younger and older age groups (Kaufman & Kaufman, 1983b, p. 104).

Concurrent Validity
The K-ABC Achievement Score was related to other criterion measures of language development, though the Expressive Vocabulary subtest was not examined separately. In various samples, the K-ABC Achievement Score showed strong correlations with the verbal scales of two major criterion measures, the Wechsler Intelligence Scale for Children—Revised (WISC-R; Wechsler, 1991) and the Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R; Wechsler, 1989; see Kaufman & Kaufman, 1983b, p. 113).

- *WISC-R Verbal Scale*:
  - In a sample of normally developing children (N = 182), r = .78.
  - In a sample of learning disabled children (N = 138), r = .74.
  - In a sample of children referred for learning disability screening (N = 60), r = .80.
  - In a sample of children with behavioral disorders (N = 43), r = .87.
  - In a sample of "educable mentally retarded" children (N = 69), r = .54.
  - In a sample of Sioux children (N = 40), r = .85.
  - In a sample of Navajo children (N = 33), r = .84.
- *WPPSI-R Verbal Scale:*
  - In a sample of normally developing preschool children (N = 40), r = .64.

Predictive Validity
The K-ABC Achievement Scores were highly predictive of measures of cognitive development. However, the Expressive Vocabulary subtest was not examined separately. Correlations of the K-ABC Achievement Score are as follows: (See Kaufman & Kaufman, 1983b, p. 121).

- With 11 months between test administrations, the correlation of the K-ABC Achievement Score and the PIAT (Markwardt, 1998) total score in a sample of normally developing children (N = 29) was .72.
- With 10 months between test administrations, the correlation between the K-ABC Achievement Score and the PIAT total score in a sample of Navajo children (N = 30) was .82.

- With 7 months between test administrations, the correlation of the K-ABC Achievement Score with the PIAT total score in a sample of "educable mentally retarded" children (N = 29) was .67.
- With 11 months between test administrations, the correlation of the K-ABC Achievement Score with the Woodcock-Johnson (WJ-R; Woodcock & Johnson, 1989) preschool cluster score in a sample of normally developing children (N = 31) was .73.
- With 6 months between test administrations, the correlation of the K-ABC Achievement Score with the Iowa Test of Basic Skills Vocabulary subtest score in a sample of normally developing children (N = 18) was .88.
- With 12 months between test administrations, the correlation of the K-ABC Achievement Score with the California Achievement Test (CAT5; CTB/McGraw Hill, 1992) Total Language Battery score in a sample of normally developing children (N = 45) was .69.

Comments

- In regard to concurrent validity, correlations were strong across various groups of children and multiple criterion measures, illustrating the acceptable validity of the K-ABC Achievement Scale as one that addresses vocabulary. It should be noted that presented correlations were for not for the Expressive Vocabulary subtest alone, but were of the Achievement Scale to which that Expressive Vocabulary subtest is one of six subtests included.
- As asserted by the authors, it is concurred that the K-ABC is highly predictive of later cognitive ability and achievement. This is seen over multiple measures and various lengths of times between assessments. The authors mentioned the caveats that correlations to the PIAT criterion might underestimate the true relationship, due to the range restriction of PIAT scoring, whereas relationships to the WJ-R could be inflated because of the heterogeneity of the K-ABC standard scores.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- In a sample of mostly black Head Start Children, the K-ABC Achievement score was found to be highly related to a measure of receptive vocabulary, the PPVT-R, though the K-ABC subtest for Expressive Vocabulary was found less (but still strongly) related (Bing & Bing, 1985).

- There are a great many studies that assess K-ABC reliability, validity, and the degree to which it can be generalized to other populations. A full list of references can be found at www.agsnet.com. For further reviews of the reliability and validity of this measure see Anatassi (1984) and Das (1984).

Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- Burchinal, Peisner-Feinberg, Bryant, and Clifford (2000) used the K-ABC Achievement Score as a language outcome in a study of child care quality. While quality (as measured using the ECERS) predicted language outcomes for all racial/ethnic groups in the sample,

it was a stronger predictor for children of ethnic minority backgrounds, even after controlling for SES and gender.

Comments

- Validity information is lacking for the Expressive Vocabulary subtest considered separately.

V. Adaptations of Measure

None found.

Early Childhood Measures: Language

Early Childhood Workshop: Language

Peabody Picture Vocabulary Test—Third Edition (PPVT-III)

I. Background Information

Author/Source

Source: Dunn, L., & Dunn, L. (1997). *Peabody Picture Vocabulary Test—Third Edition: Examiner's Manual*. Circle Pines, MI: American Guidance System.

Publisher: American Guidance Systems
4201 Woodland Road
Circle Pines, MN 55014
Phone: 800-328-2560
www.agsnet.com

Purpose of Measure

As described by instrument publisher:
"The PPVT-III is a test of listening comprehension for the spoken word in standard English. It has two purposes: First, the PPVT-III is designed as a measure of an examinee's receptive (hearing) vocabulary. In this sense it is an achievement test of the level of a person's vocabulary acquisition. Second, the PPVT-III serves as a screening test for verbal ability, or as an element in a comprehensive battery of cognitive processes. However, it can be used for this second purpose only when English is the language spoke in the examinee's home, community, and school" (Dunn & Dunn, 1997, p.2).

Population Measure Developed With

- The norming sample included 2,725 participants, ranging in age from 2 years, 6 months to 90+ years.
- Sampling was done so that the standardization population roughly matched the general U.S. population (1994) for age, sex, geographic location, parental education level (or if an adult was being tested, own education level), and race/ethnicity (black, Hispanic, white, other).
- The sample distribution was also matched to the current population for special needs groups: learning disabled, speech impaired, mentally retarded, hearing impaired, as well as gifted and talented.
- Because of the rapid language development of children from the ages of 2 years, 6 months to 6 years, children in this age range were divided into 6-month age intervals.
- For ages 7 to 16, a period with a steady but slower increase in vocabulary, whole year intervals were used.
- For the adult ages, where vocabulary growth rate slows further and eventually begins descending, multi-year intervals were used.

Age Range Intended For

Ages 2 years, 6 months through 90+ years.

Key Constructs of Measure

- Receptive language ability for standard English.

Norming of Measure (Criterion or Norm Referenced)

Norm referenced.

Comments

- It is noteworthy that the norming sample was diverse in terms of sociodemographic characteristics, and included special needs groups.
- This measure is appropriate for a wide age range.
- Concerns were raised about possible cultural bias for earlier versions of the PPVT, especially the possibility that it underestimated the abilities of minority children. At the same time, research indicated that the PPVT-R predicted IQ scores for black children as well as white children (Halpin, Simpson and Martin, 1990), and predicted IQ and achievement scores for at-risk preschoolers (Bracken and Prasse, 1983; Kutsick, Vance, Schwarting and West, 1988).

II. Administration of Measure

Who is the Respondent to the Measure?

Child (age range extends into adulthood).

If Child is Respondent, What is Child Asked to Do?

Children are presented with Picture Plates. Each Picture Plate presents four numbered cards simultaneously. Only one card represents a stimulus word pictorially. The children are asked to identify verbally or behaviorally which card represents the stimulus word (e.g. if a pointing response, "Put you're finger on *digging*." If a verbal response, "What number is *digging*?"

Who Administers Measure/Training Required?

Test Administration:

- Formal training in psychometrics is not required to administer this assessment, especially with populations that are generally "easy-to-test."
- The examiner should be thoroughly familiar with the test materials and instruction manual.
- He/she should also practice administering the test and using the scoring materials, preferably under the scrutiny of a trained examiner.

Data Interpretation:

- Interpretation requires a background in psychological testing and statistics.

Setting (e.g., one-on-one, group, etc.):

One-on-one.

Time Needed and Cost

Time:

- 11 to 12 minutes

Cost:

- Basic test: $262.95
- Test with accompanying computer package: $361.95

III. Functioning of Measure

Reliability Information from the Manual

Alternate-Forms Reliability

Alternate forms of the same test were given in a counterbalanced design. Alternate-forms reliability coefficients for standard scores ranged from .88 to .96 (median = .94) and coefficients for raw scores ranged from .89 to .99 (median = .95) (Dunn, & Dunn, 1997, p. 49).

Internal Reliability

Alpha coefficients ranged from .92 to .98 (median = .95), varying by age (Dunn, & Dunn, 1997, p. 50).

Split-Half Reliability

Split-half reliability ranged from .86 to .97 (median = .94; Dunn, & Dunn, 1997, p. 50).

Validity Information from the Manual

Internal Validity

Goodness-of-fit statistics were used to establish the degree to which test items matched established growth curves for language development. The authors reported that item response reasonably matched these growth curves to establish that items are placed in correct order of difficulty.

Criterion Validity

Correlations between the PPVT-III (Form A and Form B, respectively) ranged from .62 to .91. The ages of the various samples differed by the criterion measures being assessed, and correlations are as follows (Dunn, & Dunn, 1997, p. 58):

- *Wechsler Intelligence Scale for Children—Third Edition* (WISC-III; Wechsler, 1991):
  - Sample age range 8 to 14.
  - Verbal IQ: r = .91; .92.
  - Performance IQ: r = .82; .84.
  - Full Scale IQ: r = .90; .90
- *Kaufman Adolescent and Adult Intelligence Test* (KAIT; Kaufman & Kaufman, 1993):
  - Sample age range 13 to 18.
  - Crystallized IQ: r = .87; .91.
  - Fluid IQ: r = .76; .85.
  - Composite IQ: r = .85; .91.

- *Kaufman Brief Intelligence Test* (K-BIT; Kaufman & Kaufman,1990):
  - Sample age range 18 to 71.
  - Vocabulary: r = .82, .80.
  - Matrices: r = .65, .62.
  - Composite: r = .78, .76.
- *Oral and Written Language Scales* (OWLS; Carrow-Woolfolk, 1995):
  - Sample age range 3 to 6 and 8 to 12.
  - LC: r = .70; .77.
  - OE: r = .67; .68.
  - Oral Composite: r = .75; .77.

Comments

- Correlations between the PPVT-III and the four chosen measures of cognitive development were generally strong, showing closer relationships between the PPVT-III and the verbal scales of the criterion measure (when applicable). This suggests that while receptive language and cognitive ability are highly related, PPVT-III scores do seem to be associated with abilities specifically related to language, thus supporting the validity of the measure.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

Note: Results detailed here are based on the PPVT-R rather than the PPVT-III.

- McCartney (1984) found a relationship between child care quality in center care (as measured with the ECERS) and language ability in preschool children using the PPVT-R, net of family background variables, age of entry into center care, and number of hours in current center care.

- Peisner-Feinberg and Burchinal (1997) examined language development in relation to concurrent child care quality in the Cost, Quality and Outcomes Study. Children's scores on the PPVT-R were significantly related to child care quality (measured through a composite rating of observed classroom quality and a rating of the teacher-child relationship) after adjusting for child and family characteristics.

Comments

- Due to the relatively recent publication date of the PPVT-III, very few studies have used this version of the assessment. However, an earlier version of the measure (PPVT-R), has been found to be highly correlated with scale scores on the Stanford-Binet. When comparisons of classificatory accuracy were made between the two (the latter as the criterion measure), it was found that the PPVT-R misclassified 45 percent of children from 2 to 15 by +/- one level of functioning, and 11 percent of children by +/- two classification levels. This underscores the suggestion in the Manual to use the PPVT as part of an assessment battery and not as a sole assessment (Tarnowski & Kelly, 1987).

- Bracken (1987) criticized the PPVT-R for having low alternate-form reliability (ranged from .76 to .79). As noted above, the PPVT-III exceeds this range for alternate-form reliability.
- Reading is not required of the examinees, nor are they required to use oral or written responses to the stimulus questions. This may allow the PPVT-III to be more easily adapted to special populations, for instance, the hearing impaired or those with speech pathologies.
- The PPVT-III stimulus illustrations have been updated to provide better ethnic and gender balance.
- The administration of the test is relatively simple and could possibly be done by someone at a para-professional/assistant level (with appropriate training), though interpretation of the data requires someone with more experience.

V. Adaptations of Measure

Spanish Version of PPVT-III:
A Spanish version of the PPVT-III is available.

Early Childhood Measures: Language

Early Childhood Measures: Language

Sequenced Inventory of Communication Development—Revised (SICD-R)

I. Background Information

Author/Source

Source: Hedrick, D., Prather, E., & Tobin, A., (1984). *Sequenced Inventory of Communication Development—Revised Edition: Test Manual*. Los Angeles, CA: Western Psychological Services.

Publisher: Western Psychological Services
12031 Wilshire Boulevard
Los Angeles, CA 90025-1251
Phone: 800-648-8857
Website: www.wpspublish.com

Purpose of Measure

As described by instrument publisher:
"The SICD-R is in fact a screening tool of communications, although not in the sense of a quick, easy procedure to separate potential problems from normal behaviors. Instead, it 'screens' broad spectrums of behavior suggesting further areas of in-depth assessment. This test is extremely useful in suggesting recommendations for the initiation of remedial programming. The clinician's ability to see patterns of strength and weakness from SICD-R profiles allows her or him to establish general, long-term objectives necessary for the child's Individualized Education Program" (Hedrick, Prather, & Tobin, 1984, p. 7).

Population Measure Developed With

- The standardization sample included 252 children who ranged from 4 months to 4 years in age, with 3 age subgroups within each year (e.g., within those who were between the ages of 0 and 1, there was a 4-month-old group, an 8-month-old group, and a 12-month-old group).
- Each age subgroup included 21 children and evenly represented three "social class" groups (i.e. high, medium, low) determined by parent education and occupation.
- Only white children were used in the sample, and approximately half of the children in each age subgroup were boys and girls (though the ratio was not exact).
- A child was excluded if his or her language development was deemed abnormal by the parent, or if the child came from a home where a language other than English was spoken.
- A child was excluded if he/she did not have hearing in the normal range or displayed physical or mental abnormalities that were obvious to the examiner.

Age Range Intended For

Ages 4 months through 4 years.

Key Constructs of Measure

The SICD-R measures two main constructs, each with several parts:

- The Receptive Language scale is comprised of three main components:
  - Awareness: The degree to which the child is observed to respond to speech, and to sounds other than human vocalization.
  - Discrimination: Parent report of the degree to which the child responds to speech and speech cues differentially.
  - Understanding: The degree to which the child is observed to respond to verbally directed tasks (broken down into speech with situational cues and speech without cues).
    - All items that test these behaviors are sequenced according to the chronological age at which 75 percent or more of children exhibit them.
    - Receptive language is broken down into semantic, syntactic, pragmatic, and perceptual content.
- The Expressive Language scale includes five factors; three reflect communicative behavior (Initiating, Imitating, and Responding) and two reflect linguistic behavior (Verbal Output, and Articulation). As in the receptive scale, direct assessment is supplemented by parent observation.

Norming of Measure (Criterion or Norm Referenced)

Norm referenced.

Comments

- The standardization sample consisted only of white children. However, the manual presents results from a further field study including a sample of black as well as white children from Detroit (in Appendix A of manual, Allen and Bliss; Detroit Field Study).
- There was a fairly small number of children within each age subgroup.

II. Administration of Measure

Who is the Respondent to the Measure?

Child. As noted, some aspects are rated through parent observation of behavior in the child's home (e.g. when asked to sit at the table for dinner, child responds by doing so).

If Child is Respondent, What is Child Asked to Do?

A range of responses is required of the child, including verbal response to questions (e.g., What do you drink from?), as well as behavioral responses to prompts. Some behavioral responses may require simple actions such as pointing (e.g., Can you point to your ear?), while others are more complex (e.g., Can you take the man from inside the car and bring him to where he sleeps?).

Who Administers Measure/Training Required?

Test Administration*:*

- The administrator of the SICD-R should be trained in its use and have background in developmental screening.

- Parents make observations of the child's behavior at home, but this does not require any specific training.

Data Interpretation*:*

- The SICDR is used to determine level of functioning and patterns of language disorder. Proper interpretation requires some knowledge of language development, patterns of functioning, and standardized assessment.

Setting (e.g., one-on-one, group, etc.)

One-on-on (assessment with trained professional).
One-on-on (parent observation in naturalistic context of the home).

Time Needed and Cost

Time:

- Ranges from 30 minutes with infants to75 minutes for children 24 months and older.

Cost:

- Complete kit: $390.00
- Manual: $32.50

Comments

- Both administration and interpretation require fairly extensive training, and the manual directions are fairly complex (e.g., for establishing ceilings and basals). Understanding scoring may require some understanding of statistics.

III. Functioning of Measure

Reliability Information from Manual

Interrater Reliability

Sixteen subjects, two or three from six of the age groups, were randomly selected to be assessed by two examiners. The mean percent of examiner agreement on items being classified as pass or fail ranged from 90.3 to 100 percent (Hedrick, Prather, & Tobin, 1984, p. 45). There is no indication in the manual as to whether there was a relationship between examiner agreement and child age group.

Reliability of Receptive Communication Age (RCA) and Expressive Communication Age (ECA) Assignments

The SCID-R provides scores for Receptive and Expressive Communication Age. The authors made RCA and ECA assignments for 21 subjects in the 32-month age subgroup. This age subgroup was selected because of the variability in performance at this age. The authors were in agreement 90.48 percent of the time (Hedrick, Prather, & Tobin, 1984, p. 46).

Test-Retest Reliability
Ten subjects were randomly selected from the 6 age groups that were not sampled from for the assessment of interrater reliability and were given the test again by the same examiner a week after the original. The mean percent agreement across time points was 92.8 percent and ranged from .88 to 98.6 percent. As a whole, the subjects did better at the second testing, perhaps due to familiarity of the test or increased comfort with the test setting. This tendency increased with child age (Hedrick, Prather, & Tobin, 1984, p. 45).

Validity Information in the Manual

Construct/Concurrent Validity

- Correlations between the RCA and ECA placements of the SICD-R and the following subscales or other tests of language development ranged from .74 to .95. Though intercorrelations of scales of the same measure are generally considered as evidence of construct validity, and correlations with a separate measure as concurrent validity, the designation is not made by the authors. The two scale scores of the SICD-R were the most highly correlated, with the criterion measures less so. According to the authors, the correlations with the other language assessments show discrimination between SICD-R constructs when compared with the Peabody Picture Vocabulary Test (PPVT; Dunn, 1965), a measurement of receptive language, as RCA correlations are slightly stronger with this criterion measure. Correlations are as follows (Hedrick, Prather, & Tobin, 1984, p. 46):
    - Receptive Communication Age and Expressive Communication Age: r = .95.
    - Receptive Communication Age and the Peabody Picture Vocabulary Test: r = .81.
    - Expressive Communication Age and the Peabody Picture Vocabulary Test: r = .76.
    - Expressive Communication Age and Mean Length of Response: r = .76.
    - Expressive Communication Age and Structural Complexity Score: r = .74

Discriminant Validity
It is noted in the manual that in previous studies, children with Down's Syndrome, autism, hearing loss, and multiple handicaps have scored significantly below their more "normally developing" peers when using the SICD- R. However, no quantitative data are provided.

Comments

- The authors do not provide information regarding the reliability of the parent report components of the assessment, and reliability might be different for parents.
- The authors' assertions regarding the divergence of correlation strength between the SICD-R scales and the scales of other comparable, yet not identical, measures of language is somewhat substantiated. That is, correlations were lower for less comparable constructs.
- All reported reliability results were strong.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- Child care quality was predictive of positive language development (as assessed by the SICD-R) for a sample of 79 black, one-year-old children (Burchinal, Roberts, Nabors & Bryant, 1996).

- In a sample of 89 low-income, black children followed longitudinally, better child care quality predicted higher language development scores as assessed by the SICD-R at 12, 24 and 36 months, net of child and family factors (Burchinal, Roberts, Riggins, Zeisel, Neebem & Bryant, 2000).

- In the Colorado Adoption Project, home environment scores (assessed using the HOME Inventory) were positively related to language development as assessed using the SICD-R for those raised by adoptive families as well as those raised by shared gene-environment families (Thompson, Fulker, DeFries & Plomin, 1986).

Comments

- The SICD assesses a variety of early communication skills.

- Although adaptations to the measure have been made to meet the needs of special populations, the psychometrics for these adaptations are not reported on.

- It would be helpful to see further validity information, such as predictive and construct validity, given the very strong relationship between the two major constructs assessed by this measure.

V. Adaptations of Measure

Yup'ik Sequenced Inventory of Communication Development

Description of Adaptation:

SICD adapted for use with Yup'ik Eskimos.

Psychometrics of Adaptation:
Not currently available.

Study Using Adaptation:
See Prather, E., Reed, C., Foley, L., Somes, & Mohr, R. (1979). *Yup'ik Sequenced Inventory of Communication Development*, Anchorage Rural Alaska Community Action Program, Inc.

SICD for Autistic and "Difficult-to-Test" Children

Description of Adaptation:

SICD adapted for use with autistic and "difficult-to-test" children.

Psychometrics of Adaptation:
Not currently available.

Study Using Adaptation:
See O'Reilly, R. (1981). *Language testing with children considered difficult to test* (Masters Thesis). Arizona State University. See also, Tominac, C. (1981). *The effect of intoned versus neutral stimuli with autistic children* (Masters Thesis). Arizona State University.

SICD for Hearing-Impaired Children

Description of Adaptation:
SICD adapted for use with hearing-impaired children.

Psychometrics of Adaptation:
Not currently available.

Study Using Adaptation:
See Oystercamp, K. (1983). *Performance of hearing-impaired children on developmental language tests* (Masters Thesis). Arizona State University.

Early Childhood Measures: Language

Early Childhood Measures: Language

Test of Early Language Development—Third Edition (TELD-3)

I. Background Information

Author/Source

Author: Hresko, W., Reid, D., & Hammill, D. (1999). *Test of Early Language Development – Third Edition: Examiner's Manual*. Austin, TX: PRO-ED, Inc.

Publisher: PRO-ED, Inc.
8700 Shoal Creek Boulevard
Austin, TX 78757
Phone: 800-897-3202
Website: www.proedinc.com

Purpose of Measure

As described by instrument publisher:
"The TELD-3 has 5 purposes: 1.) to identify those children that are significantly below their peers in early language development and thus may be candidates for early intervention; 2.) to identify strengths and weaknesses of individual children; 3.) to document children's progress as a consequence of early language intervention programs; 4.) to serve as a measure in research studying language development in young children; 5.) to accompany other assessment techniques" (Hresko, Reid, & Hammill, 1999, p. 7).

Population Measure Developed With

- The norming sample was selected based on geographic location, gender, race, ethnicity, family income, parental education, disability, and age.
- The sample of 2,217 children ranged from 2 to 7 years of age. There were 226 2-year-olds, 266 3-year-olds, 423 4-year-olds, 494 5-year-olds, 430 6-year-olds, and 378 7-year-olds.
- Demographics for the sample children were broadly comparable to 1990 U.S. Census data, with small differences in percentages for black children (13 percent vs. 16 percent), children whose parents obtained less than a Bachelor's degree (72 percent vs. 76 percent), and a slight overrepresentation of children whose parents had Bachelor's degrees (20 percent vs. 16 percent). The sample demographics were comparable to projected 2000 Census data.

Age Range Intended For

Ages 2 years through 7 years, 11 months.

Key Constructs of Measure

- *Receptive Language Subtest*: The Receptive Language subtest measures the comprehension of language. Children who are less successful on this subtest may have difficulty understanding directions in class, interpreting the needs of others, and understanding complex conversations.
- *Expressive Language Subtest*: The Expressive Language subtest measures the ability to communicate orally. Children who do well on this subtest should be able to answer questions, participate in conversations, use varied vocabulary, and generate complex sentences.
- *Spoken Language Quotient*: The Spoken language Quotient is a composite score based on both the Receptive and Expressive language subtests. As such, it is the best indicator of a child's overall oral language ability. Children who do poorly on this composite may have problems communicating effectively and understanding language in the home, school, and community contexts; may show difficulty in reading and writing; and may have problems engaging in social settings. Establishing the cause of such language deficits is beyond the scope of the measure.

Norming of Measure (Criterion or Norm Referenced)

Norm referenced.

Comments

- The size and diversity of the norming sample seem strong.

II. Administration of Measure

Who is the Respondent to the Measure?

Child.

If Child is Respondent, What is Child Asked to Do?

Receptive Language Subtest:

- For this subtest, the individual child is asked to exhibit a range of behavioral responses to questions and requests. For instance, examiners observe whether the child responds properly to his/her name being called, or whether the child can follow simple directions such as, "Can you sit down in the chair?" The child may be asked to point to a picture illustrating a particular word (e.g., "Can you show me the dog?") or to point to a body part (e.g., "Can you point to your nose?"). The child may be asked to use a toy in responding (e.g., "Can you put the boy on the table?"). The questions and requests become increasingly difficult. At the most difficult level, questions elicit complex understanding of words in formats like "What goes with quickly – slowly or rapidly?" or "Is a gangster a criminal?"

Expressive Language Subtest:

- Like the receptive language subtest, the expressive language subtest requires a range of responses from the individual child; responses vary as difficulty increases. For example, initial behavioral responses to be noted by the examiner include whether the child

expresses pleasure or anger. The child may be presented with illustrations of common objects (such as a tree) and asked "What is this?" The examiner records information about the complexity of the child's sentences; for example, whether the child routinely uses sentences of more than two to three words, or how many sentences the child uses to answer a question like, "Tell me about your favorite game." In addition, the examiner records whether the child uses pronouns properly when asked "Who does this toy belong to?" At the most difficult level, the child is asked questions that require greater detail to answer, such as, "Why does your father go to work?"

Who Administers Measure/Training Required?

Test Administration:

- Those who administer this test should have some formal training in administering assessments and interpreting assessment data.
- Supervised practice in using and scoring language assessments is also desirable.

Data Interpretation

- Same as above.

Setting (e.g. one-on-one, group, etc)

One-on-one.

Time Needed and Cost

Time:

- 15-40 minutes, depending on age and ability. The test is not timed.

Cost:

- Manual: $74.00
- Complete Kit: $264.00

Comments

- The complexity of this measure requires administration by trained individuals. Administration/coding/interpretation of this measure by a less-qualified individual would raise concerns of reliability and standardized interpretation of the data. This is especially a concern when the more developmentally advanced questions are asked of the child.
- For the more difficult questions of the expressive language subtest, the rater must be careful to distinguish sentence complexity from content of the response.
- This measure does not provide alternate testing methods for children with auditory, oral, or physical impairments.

III. Functioning of Measure

Reliability Information from Manual

Internal Reliability

Internal reliability coefficients were high for each of the TELD-3 subtests, with mean coefficient alphas (across age groups) of .91, .92, and .95 for the Receptive, Expressive, and Spoken

Language Subtests, respectively. There was very little variation with child age, with alphas for the oldest age group slightly lower than for the other age groups (.80-.89; Hresko, Reid, & Hammill, 1999, p. 81).

Test-Retest Reliability
Zero-order correlations were calculated between test scores taken two weeks apart. Correlations differed somewhat by age, ranging from .83 to .87 for Receptive Language and from .82 to .93 for Expressive Language (Hresko, Reid, & Hammill, 1999, p. 85). For the Receptive and Expressive subtests, test-retest correlations were highest for the youngest children (ages 2 to 4) and lower for oldest children (ages 5 to 7).

Validity Information from Manual
Criterion-Prediction Validity
Correlations between the two subtests and the composite score were given for each of the two forms of the measure (Form A and Form B) and various other tests of language development. Correlations ranged from .30 to .78 except for the correlations with a previous edition of the measure (TELD-2; Hresko, Reid, & Hammill, 1991), which were much higher (.84 to .92). The correlations are presented below. The first correlation coefficient represents Form A of the measure, and the second correlation coefficient represents Form B (Hresko, Reid, & Hammill, 1999, p. 104):

- *Expressive Language*:
  - Communication Abilities Diagnostic Test (Johnston & Johnston, 1990): r = .48; .40.
  - Clinical Evaluation of Language Fundamentals—Preschool (Wiig, Secord, & Semel, 1992): r = .59; .65.
  - Expressive One-Word Vocabulary Test (Brownell, 2000); r = .44; .30.
  - TELD-2 (Hresko, Reid, & Hammill, 1991): r = .87; .84.
  - Peabody Picture Vocabulary Test (Dunn, 1965); r = .73; .83.
  - Preschool Language Scale–3 (Zimmerman, Steiner, & Pond, 1992): r = .70; .74.
  - Receptive One-Word Vocabulary Test (Brownell, 2000): r = .53; .40.
- *Receptive Language*:
  - Communication Abilities Diagnostic Test: r = .40; .40.
  - Clinical Evaluation of Language Fundamentals- Preschool: r = .55; .71.
  - Expressive One Word Vocabulary Test; r = .41; .44.
  - TELD-2: r = .84; .84.
  - Peabody Picture Vocabulary Test; r = .67; .70.
  - Preschool Language Scale –3: r = .55; .62.
  - Receptive One-Word Vocabulary Test: r = .40; .40.
- *Spoken Language*:
  - Communication Abilities Diagnostic Test: r = .45; .44.
  - Clinical Evaluation of Language Fundamentals—Preschool: r = .77; .76.
  - Expressive One Word Vocabulary Test; r = .42; .38.
  - TELD-2: r = .90; .92.
  - Peabody Picture Vocabulary Test; r = .79; .84.
  - Preschool Language Scale –3: r = .61; .70.

- Receptive One-Word Vocabulary Test: r = .50; .48.

Relation to Intelligence
Relationships between the TELD-3 and selected intelligence test scores (TELD-3 Form A, then Form B) are presented below (Hresko, Reid, & Hammill, 1999, p. 110):

- *Receptive Language*:
  - Stanford-Binet Intelligence Scales-IV (Thorndike, Hagen, & Sattler, 1986): r = .41; .41.
  - Wechsler Intelligence Scales for Children-III (Wechsler, 1991): r = .62 (verbal), .44 (performance), .47 (full); .65 (verbal), .66 (performance), .48 (full).
  - Woodcock-Johnson Psychoeducational Battery—Revised (Woodcock, & Johnson, 1989): r = .55; .72.
- *Expressive Language*:
  - Stanford-Binet Intelligence Scales-IV: r = .46; .46.
  - Wechsler Intelligence Scales for Children-III: r = .57 (verbal), .43 (performance), .47 (full); .73(verbal), .63 (performance), .71 (full).
  - Woodcock-Johnson Psychoeducational Battery—Revised: r = .56; .59.
- *Spoken Language:*
  - Stanford-Binet Intelligence Scales-IV: r = .43; .46.
  - Wechsler Intelligence Scales for Children-III: r = .67 (verbal), .52 (performance), .67 (full); .76 (verbal), .65 (performance), .64 (full).
  - Woodcock-Johnson Psychoeducational Battery—Revised; r = .64; .64.

Relation to Age
Scores on both the Receptive Language scale and the Expressive Language scale improved with age, and correlations over both forms of each scale and age ranged from .80 to .86, with Expressive Language being somewhat less related than Receptive.

Comments

- For the criterion-predictive validity, moderate to high correlations indicate that while there is relationship between the TELD-3 constructs and the criterion measures, the TELD-3 (or criterion) may be capturing a different part of the construct than the criterion to which it is being compared. As stipulated by the authors, these moderate to high correlations could be argued to show good validity of the measure. The stronger relationship between the TELD-3 and the former version, which was based on different sample and has spanned time, suggests acceptable validity for the current version. Though correlations are in the moderate to high range, correlations between the specific Expressive and Receptive subtests of the TELD-3 do not greatly differentiate between criterions that measure either expressive or receptive language, specifically. That is, correlations are not higher between the TELD-3 Receptive scale and the Receptive One-Word Vocabulary Test than the Expressive One Word Vocabulary Test.
- As seen in the comparisons to other vocabulary tests, correlations between the TELD-3 scales and tests of cognitive ability are generally within the moderate to high range. Some shared variance is expected between intelligence and language, but similar to the comparison of the TELD-3 and the language specific criterions, a moderate relationship may suggest that the TELD-3 (or the criterion) measures aspects of its constructs beyond

intelligence alone. The authors of the measure assert that a significant correlation between the two supports the validity of the measure.
- Though correlational trends exist between age and TELD-3 scores, age group mean scores were not individually tested between each other, thus the significance of the difference between each age group is unknown.

Reliability/Validity Information from Other Studies

- Mcloughlin and Gullo (1984) assessed the predictive validity of the PPVT-R (Dunn & Dunn, 1981) and the TELD using the Preschool Language Scale (PLS) as the criterion measure. Together, the PPVT-R and the TELD accounted for less than 47 percent of the variance in the PLS. When unique variance was assessed, it was found that the unique predictive value of the TELD dropped below 1 percent after the PPVT-R was taken into account. However, this was not the case for the PPVT-R. When the variance accounted for by the TELD was partialed, the PPVT-R remained a significant predictor of PLS scores. It should be noted that this study used the first version of the TELD, and results addressed might not apply to the TELD-3 version.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

None found.

Comments

- It is unclear why test-retest reliability decreases somewhat with age.

- The TELD-3 provides scores for both expressive and receptive language as well as a global rating of language development.

V. Adaptations of Measure

None found.

Language and Literacy Measure References

Anastasi, A. (1984). The K-ABC in historical and contemporary perspective. *Journal of Special Education*, *18*(3), 357-366.

Arriaga, R. I., Fenson, L., Cronan, T., & Pethick, S. J. (1988). Scores on the MacArthur Communicative Development Inventory of Children from low and middle-income families. *Applied Psycholinguistics, 19,* 209-223.

Bing, S., & Bing, J. (1985). Comparison of the K-ABC and PPVT-R with Head Start children. *Psychology in the Schools, 22*, 245-249.

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

Bracken, B.A., & Prasse, D. P. (1983). Concurrent validity of the PPVT-R for "at –risk" preschool children. *Psychology in the Schools, 20*(1)*,* 13-15.

Brownell, R. (2000). *Expressive One-Word Picture Vocabulary Test: Manual*. Novato, CA: Academic Therapy Publications.

Brownell, R. (2000). *Receptive One-Word Picture Vocabulary Test: Manual*. Novato, CA: Academic Therapy Publications.

Burchinal, M., Peisner-Feinberg, E., Bryant, D., & Clifford, R. (2000). Children's social and cognitive development and child-care quality: Testing differential associations related to poverty, gender, or ethnicity. *Applied Developmental Science, 4(3)*, 149-165.

Burchinal, M., Roberts, J., Riggins, R., Zeisel, S., Neebe, E., & Bryant, D. (2000). Relating quality of center-based child care to early cognitive and language development longitudinally. *Child Development, 71(2),* 339-357.

Burchinal, M. R., Roberts, J. E., Nabors, L. A., & Bryant, D.M. (1996). Quality of center child care and infant cognitive and language development. *Child Development, 67*, 606–620.

Carrow-Woofolk, E. (1985). Test for Auditory Comprehension of Language- Revised Edition. Austin, TX: Pro-Ed.

Carrow-Woofolk, E. (1995). Oral and Written Language Scales. Circle Pines, MN: American Guidance Service.

CTB/McGraw Hill. (1992). *California Achievement Tests, 5th ed.* Monterey, CA: Author.

Das, J. P. (1984). Review of the Kaufman Assessment Battery for Children. *Journal of Psychoeducational Assessment, 2*, 49-56.

Dunn, L.M. (1965). *Peabody Picture Vocabulary Test: Expanded manual*. Circle Pines, MD: American Guidance Service.

Dunn, L., & Dunn, L. (1981). *Peabody Picture Vocabulary Test – Revised*. Circle Pines, MN: American Guidance Service.

Dunn, L., & Dunn, L. (1997). *Peabody Picture Vocabulary Test- Third Edition: Examiner's manual.* Circle Pines, MN: American Guidance Systems.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development, 71(2),* 310-322.

Fenson, L., Bates, E., Dale, P., Goodman, J., Reznick, J., & Thal, D. (2000). Measuring variability in early childhood language: Don't shoot the messenger. *Child Development, 71*(2), 323-328.

Fenson, L., Dale, P., Reznick, S., Thal, D., Bates, E., Hartung, J., et al. (1993) *MacArthur Communicative Development Inventories: User's guide and technical manual.* San Diego, CA: Singular/ Thomson Learning.

Goelman, H., & Pence, A. (1987). Some aspects of the relationships between family structure and child language development in three types of day care. *Advances in Applied Developmental Psychology, 2,* 129-146.

Halpin, G., & Simpson, R.G. (1990). An investigation of racial bias in the Peabody Picture Vocabulary Test Revised. *Educational and Psychological Measurement, 50*(1), 183 – 189.

Harcourt Brace Educational Measurement. (1992). Metropolitan Achievement Test, 7th ed. San Antonio, TX: Author.

Harcourt Brace Educational Measurement. (1996). Stanford Achievement Test, 9th ed. San Antonio, TX: Author.

Hedrick, D., Prather, E., & Tobin, A., (1984). *Sequenced Inventory of Communication Development-Revised Edition*: *Test manual*. Los Angeles, CA: Western Psychological Services.

Hresko, W.P., Reid, D.K., & Hammill, D.D. (1991). *Test of Early Language Development-Second Edition*. Austin, TX: PRO-ED.

Hresko, W., Reid, D., & Hammill, D. (1999). *Test of Early Language Development – Third Edition: Examiner's manual*. Austin, TX: PRO-ED, Inc.

Ingram, D. (1981). *Procedures for the phonological analyses of children's language.* Blatimore, MD: University Park Press.

Johnston, E.B., & Johnston, A.V. (1990). *Communicative Abilities Diagnostic Test*. Chicago: Riverside.

Kaufman, A. S., & Kaufman, N. L. (1983a). *Kaufman Assessment Battery for Children: Administration and scoring manual.* Circle Pines, MN: American Guidance Service, Inc.

Kaufman, A. S., & Kaufman, N. L. (1983b). *Kaufman Assessment Battery for Children: Interpretive manual.* Circle Pines, MN: American Guidance Service, Inc.

Kaufman, A. S., & Kaufman, N. L (1990). *Kaufman Brief Intelligence Test.* Circle Pines, MN: American Guidance Services.

Kaufman, A. S., & Kaufman, N. L (1993). *Kaufman Adolescent and Adult Intelligence Test.* Circle Pines, MN: American Guidance Service.

Kutsick, K., Vance, B., Schwarting, F.G., and West, R. (1988). A comparison of three different measures of intelligence with preschool children identified at risk. *Psychology in the Schools, 25*(3)*,* 270-275.

Markwardt, F. C. (1998). *Peabody Individual Achievement Test—Revised: Manual* (Normative update). Circle Pines, MN: American Guidance Service. Princeton, NJ: Mathematica Policy Research Inc. DHHS-105-95-1936.

McCartney, K. (1984). Effect of quality of day care environment on children's language development. *Developmental Psychology, 20*(2), 244-260.

Mcloughlin, C., & Gullo, D. (1984). Comparison of three formal methods of preschool language assessment. *Language, Speech, & Hearing Services in Schools, 15(3)*, 146-153.

Miller, J.F. (1981). *Assessing language production in children.* Baltimore, MD: University Park Press.

Newcomer, P. & Hammill, D. (1997). Test of Language Development- Primary. Austin, TX: Pro-Ed.

NICHD Early Child Care Research Network (2000). The relation of child care to cognitive and language development. *Child Development, 71(4),* 960-980.

O'Reilly, R. (1981).Language *Testing with Children Considered Difficult to Test*. Master's Thesis. Arizona State University.

Osterkamp, K. (1983). *Performance of hearing impaired children on developmental language tests*. Unpublished master's thesis, Arizona State University.

Otis, A. & Lennon, R. (1982). *Otis-Lennon School Ability Test.* San Antonio, TX: Harcourt & Brace.

Pena, E., Quinn, R., & Iglesias, A. (1992). The applications of dynamic methods to language assessment: A nonbiased procedure. *The Journal of Special Education, 26(3),* 269-280.

Peisner-Feinberg, E. S., & Burchinal, M. R. (1997). Relations between preschool children's child care experiences and concurrent development: The Cost, Quality, and Outcomes Study. *Merrill-Palmer Quarterly, 43*(3), 451-477.

Pommer L. (1986). Seriously emotionally disturbed children's performance on the Kaufman Assessment Battery For Children: A concurrent validity study. *Journal of Psychoeductional Research, 4,* 155 – 162.

Prather, E., Reed, I., Foley,. C., Somes, L., & Mohr, R. (1979). *Yup 'ik sequenced inventory of communication development*. Anchorage: Rural Alaska Community Action Program, Inc.

Semel, E., Wiig, E., & Secord, W. (1995) Clinical Evaluation of Language Fundamentals, 3rd ed. San Antonio, TX: The psychological Corporation.

Tarnowski, K., & Kelly, P. (1987). Utility of PPVT for pediatric intellectual screening. *Journal of Pediatric Psychology, 12(4),* 611-614.

Thal, D. & Bates, E. (1988). Language and gesture in late talkers. *Journal of Speech and Hearing Research, 31,* 115-123..

Thompson, L., Fulker, D., DeFries, J., & Plomin, R. (1986). Multivariate genetic analysis of "environmental" influences on infant cognitive development. *British Journal of Developmental Psychology, 4(4),* 347-353.

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

Tominac, C. (1981) *The effect of intoned versus neutral stimuli with autistic children*. Unpublished masters thesis, Arizona State University.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence – Revised (WPPSI-R): Manual*. San Antonio, TX: The Psychological Corporation, Harcourt Brace Jovanovich, Inc.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children – 3rd ed.* San Antonio, TX: The Psychological Corportation.

Wechsler, D. (1991). *Wechsler Preschool and Primary Scale for Children – Third Edition.* San Antonio, TX: The psychological Corporation.

Wechsler, D. (1997).WAIS-III Administration and Scoring Manual.San Antonio: The Psychological Corporation.

Wiig, E.H., Secord, W., & Semel, E. (1992). *Clinical Evaluation of Language Fundamentals-Preschool.* San Antonio, TX: Psychological Corporation.

Williams, K. (1997). *Expressive Vocabulary Test*. Circle Pines, MN: American Guidance Service.

Woodcock, R.W. & Johnson, M.B. (1989). *Woodcock-Johnson Psycho-Educational Battery – Revised*. Itasca, IL: Riverside Publishing.

Zimmerman, I., Steiner, V. & Pond, R. (1992). Preschool Language Scales-3. San Antonio, TX: The Psychological Corporation.

Early Childhood Measures: Math

Early Childhood Measures: Math

Bracken Basic Concept Scale—Revised (BBCS-R), Math Subtests

I. Background Information

Author/Source

Source: Bracken, B. A. (1998). *Bracken Basic Concept Scale—Revised: Examiner's Manual*. San Antonio, TX: The Psychological Corporation.

Publisher: The Psychological Corporation
19500 Bulverde Rd.
San Antonio, TX 78259
Phone: 800-872-1726
Website: www.psychcorp.com

Purpose of Measure

A summary of BBCS-R is also included within the Cognitive Assessment section of this compendium. Some of the information in the two summaries is the same, but in this summary we provide more detailed information on subtests related to mathematics.

As described by instrument publisher:
This measure is designed to assess children's concept development and to determine how familiar children are with concepts that parents, preschool teachers, and kindergarten teachers teach children to prepare them for formal education.

"The BBCS-R English edition serves five basic assessment purposes: speech-language assessment, cognitive assessment, curriculum-based assessment, school readiness screening, and assessment for clinical and educational research" (Bracken, 1998, p. 6).

Population Measure Developed With

- The standardization sample was representative of the general U.S. population of children ages 2 years, 6 months through 8 years and was stratified by age, gender, race/ethnicity, region, and parent education. Demographic percentages were based on 1995 U.S. Census data.
- The sample consisted of 1,100 children between the ages of 2 years, 6 months and 8 years.
- In addition to the main sample, two clinical studies were conducted—one with 36 children who were developmentally delayed, and one with 37 children who had language disorders.

Age Range Intended For

Ages 2 years, 6 months through 8 years

Key Constructs of Measure
The BBCS-R includes a total of 308 items in 11 subtests tapping "…foundational and functionally relevant educational concepts..." (Bracken, 1998, p. 1). There are four subtests related to math:

- *Numbers/Counting*: Number recognition and counting abilities.
- *Sizes*: Understanding of one-, two-, and three-dimensional size concepts such as tall, short, and thick.
- *Shapes*: Knowledge of basic one-, two-, and three-dimensional shapes (e.g., line, square, cube), and abstract shape-related concepts (e.g. space).
- *Quantity*: Understanding of concepts involving relative quantities, such as a lot, full, and triple.

However, the first three of the math subtests are part of the School Readiness Composite (SRC), which consists of a total of six subtests (i.e., Colors, Letters, Numbers/Counting, Sizes, Comparisons, and Shapes). The Manual provides the scoring procedures and psychometric properties for the SRC, but not for its six component subtests alone. SRC subtests are not intended to be used separately. A description of the other subtests used in the SRC can be found in the BBCS-R Cognitive profile of this compendium

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

Comments

- The first six subtests make up the School Readiness Composite (SRC). In order to determine a starting point for subtests 7-11, the child must complete the full SRC. In addition, the SRC (subtests 1-6) is treated as a single subtest in the scoring guidelines provided in the Manual; subtests 1-6 are not intended to be used separately. Therefore, it might be difficult to administer or interpret the individual math subtests on their own.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
The BBCS-R is designed to minimize verbal responses. Responses are either pointing responses (i.e., the child is asked to respond by pointing to pictures) or short verbal responses. Example: "Look at all of the pictures. Show me the circle."

The BBCS-R utilizes basals and ceilings. A ceiling is established within each subtest when the child answers three consecutive items incorrectly. For the first six subtests (SRC), assessment always starts with the first item. The starting point for the rest of the subtests is determined based on the child's SRC score, and a basal is established when the child passes three consecutive items.

Who Administers Measure/Training Required?

Test Administration:

- Those who administer and interpret the results of the BBCS-R should be knowledgeable in the administration and interpretation of assessments. According to the publisher, people who are involved with psychoeducational assessment or screening (school psychologists, special education teachers, etc.) will find the test easy to administer, score, and interpret.

Data Interpretation:

- (Same as above.)

Setting (e.g., one-on-one, group, etc.)

One-on-one.

Time Needed and Cost

Time:

- The BBCS-R is untimed, so the time needed for each subtest and the full battery varies. According to Psychological Corporation's customer service, it takes about 30 minutes to administer the SRC (subtests 1 through 6).

Cost:

- Complete kit: $245
- Examiner's Manual: $63

Comments

- As noted by the publisher, because the BBCS-R minimizes verbal responses it can be used as a warm-up for other assessments. In addition, it is useful for children who are shy or hesitant, or for those with a variety of conditions that might limit participation in other assessments (e.g., social phobia, autism).

III. Functioning of Measure

Reliability Information from the Manual

Split-Half Reliability

Split-half reliability estimates were calculated by correlating total scores for odd-numbered items with total scores for even-numbered items and applying a correction formula to estimate full-test reliabilities. As in the calculations of test-retest reliability (below), analyses were conducted using the SRC (not individual tests 1 to 6) and individual tests 7 to 11. The average split-half reliabilities across ages 2 years to 7 years were .91 for the SRC and .95 for the Quantity subtest (see Bracken, 1998, p. 64).

Test-Retest Reliability

A subsample of 114 children drawn from the standardization sample took the BBCS-R twice (7-14 days apart). The subsample was drawn from three age groups—3, 5, and 7 years. As with the split-half reliability analyses, the authors did not look at subtests 1 through 6 separately, but

instead looked at the SRC scores. Analyses were conducted using the SRC and individual subtests 7 to 11, including the Quantity subtest. The test-retest reliability of the SRC was .88. The test-retest reliability of the Quantity subtest was .78 (see Bracken, 1998, p. 67).

Validity Information from the Manual

Internal Validity

Correlations were calculated for each age group (2 to 7 years), as well as for the full sample, among the SRC, subtests 7 to 11, and the full battery. Correlations between the SRC and subtests 7 to 11 for the full sample ranged from .58 (Time/Sequence) to .69 (Direction/Position). Correlations between the Quantity subtest and the SRC and other individual subtests are high, ranging from .61 (SRC) to .67 (Direction/Position and Self-/Social Awareness; see Bracken, 1998, p. 75).

Concurrent Validity

A number of studies were reported in which children's scores on the BBCS-R were correlated with scores on other measures of cognitive, language, and conceptual development, including the Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R; Wechsler, 1989), the Differential Ability Scales (DAS; Elliott, 1990), the Peabody Picture Vocabulary Test—Third Edition (PPVT-III; Dunn & Dunn, 1997), the Preschool Language Scale-3 (PLS-3; Zimmerman, Steiner, & Pond, 1992) the Boehm Test of Basic Concepts—Revised (Boehm-R; Boehm, 1986a), and the Boehm Test of Basic Concepts—Preschool Version (Boehm-Preschool; Boehm, 1986b). Across these studies, correlations between BBCS-R SRC and Total Test scores and scores on other measures were moderate to high, with most correlations falling above .70. However, none of these studies examined associations with the Quantity subtest, and none of the associations between the SRC and other subtest scores involved measures that were specifically math-related. These associations are thus more relevant to the validity of the BBCS-R as a general cognitive measure and are summarized in the BBCS-R Cognitive profile.

Predictive Validity

In a study of the predictive validity of BBCS-R over the course of a kindergarten year, BBCS-R scores, children's chronological age, social skills, and perceptual motor skills were used to predict 71 kindergarteners' academic growth, as indicated by teachers' nominations for grade retention. Demographic information for this sample was not included in the Manual. Among the variables included in this study, SRC scores and scores on subtests 7 through 11 were found to be the strongest predictors of children's academic growth (see Bracken, 1998, p. 71). Between 82 and 90 percent of children who were subsequently recommended for retention by their classroom teachers were correctly identified with SRC scores. The extent to which the Quantity subtest contributed to prediction of academic growth independent of other subtests and the SRC was not reported. Thus, as is the case for concurrent validity, these scores are relevant to the functioning of the measure as a whole, rather than of the math components.

Reliability/Validity Information from Other Studies

Since BBCS-R is a fairly recent version of the test, few studies of its psychometric properties are available, although several studies of the original BBCS—either the SRC or the assessment in its entirety—have been published. As with the concurrent validity studies reported in the Manual, however, none of these studies reported associations with the Quantity subtest, and none of the

associations between the SRC and other subtest scores involved measures that were specifically math-related. For this reason, these studies are not discussed here, but are summarized in the BBCS-R Cognitive profile.

Comments

- Information presented by Bracken (1998) for the SRC and the Quantity subtest (the one math-related subtest that was examined separately) suggests that these measures demonstrate good reliability. Reported split-half reliability estimates are high, indicating high internal consistency of these measures. Further, test-retest correlations also indicate a high degree of consistency in children's relative performance on math-related measures derived from the BBCS-R across a one- to two-week interval. As noted earlier, the SRC is a general composite, rather than an exclusively math-related measure, and the subtests that comprise the SRC are not designed to be used separately. No reliability information was provided for the three separate math-related subtests included in the SRC, and thus split-half and test-retest reliabilities of these separate subtests are unknown.
- With respect to internal validity, reported correlations among subtest, SRC, and full battery scores were high, indicating that although scores for each subtest can contribute unique information regarding children's conceptual development, there is also a substantial amount of overlap in the areas of development tapped by each subtest. Because the SRC is designed as a general composite, the only information provided by Bracken (1998) that directly addressed the internal validity of subtests tapping math-related conceptual development involves the high correlations between Quantity subtest scores and scores on other subtests.
- Although information on associations between the BBCS-R and other measures of cognitive, language, and conceptual development provides evidence of convergent validity for the BBCS-R as a whole, no information was provided that directly addresses the convergent validity of the Quantity subscale or the SRC as measures of conceptual development within the math domain. Similarly, the predictive validity of the Quantity subscale or of the math subscales included within the SRC cannot be determined from information provided by Bracken (1998).

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- The original version of BBCS was one of the measures used in the NICHD Study of Early Child Care (NICHD ECCRN, 1999). The study had a sample of 1,364 families in multiple cities. Families were recruited in 1991, and the first wave of data covered birth through 36 months of age. The BBCS was administered to children at 36 months of age, and SRC scores were used in analyses. Child care quality ratings obtained through the Observational Record of Caregiving Environment (ORCE) were not related to SRC scores. However, children whose caregivers had higher levels of education (at least some college) and training (formal, post high school) had higher scores on the SRC than did children whose caregivers had lower levels of education and training. These findings do not, however, directly address the effects of environmental variation on children's understanding of math-related concepts as distinct from overall conceptual development as assessed with the BBCS.

- The original version of the BBCS (SRC only) was used in the Child Outcomes Study of the National Evaluation of Welfare-to-Work Strategies Two Year Follow-up (McGroder, Zaslow, Moore, & LeMenestrel, 2000). This study was an experimental evaluation, examining impacts on children of their mothers' (random) assignment to a JOBS welfare-to-work program or to a control group. Two welfare-to-work program approaches (a work-first and an education-first approach) were evaluated in each of three study sites, (Atlanta, Georgia; Grand Rapids, Michigan; and Riverside, California) for a total of six JOBS programs evaluated overall in this study. Children were between the ages of 3 years and 5 years at the time of their mothers' enrollment in the evaluation, and between the ages of 5 years and 7 years at the Two Year Follow-up. The Two Year Follow-up study found an impact on SRC scores in the work-first program in the Atlanta site, with children in the program group scoring higher on the SRC than did children in the control group. This study also examined the proportion of children in the program and control groups scoring in the high and low ends of the distribution for this measure (equivalent to the top and bottom quartiles in the standardization sample). For three of the six programs, a higher proportion of children of mothers assigned to a JOBS program scored in the top quartile, compared to children of mothers in the control group. In addition, in one of the six programs, children of mothers in the program group were less likely to score in the bottom quartile on the SRC than were children of mothers in the control group. Once again, however, this study does not directly assess the impact of the program on children's understanding of math-related concepts, although it does point to a program impact on the SRC, and half of the individual subtests that comprise the SRC focus on math.

Comments

- As indicated above, no studies were found that used math subtests of the BBCS or the BBCS-R separately from SRC and full battery scores. Thus, the use of subtests of the BBCS-R as a specific measure of conceptual development within the math domain is untested.

V. Adaptation of Measure

Spanish Version

Description of Adaptation:

A Spanish version of BBCS-R is available. Spanish-language forms are designed to be used with the English-language stimulus manual. The Spanish version is to be used as a curriculum-based measure only because it is not a norm-referenced test. Field research was conducted with a sample of 193 Spanish-speaking children between the ages of 2 years, 6 months and 7 years, 11 months.

Early Childhood Measures: Math

Early Childhood Measures: Math

Kaufman Assessment Battery for Children (K-ABC), Arithmetic Subtest

I. Background Information

Author/Source

Source: Kaufman, A. S., & Kaufman, N. L. (1983a). *Kaufman Assessment Battery for Children: Administration and scoring manual*. Circle Pines, MN: American Guidance Service. (See also Kaufman, A. S., & Kaufman, N. L. (1983b). *Kaufman Assessment Battery for Children: Interpretive manual*. Circle Pines, MN: American Guidance Service.)

Publisher: American Guidance Service
4201 Woodland Road
Circle Pines, MN 55014
Phone: 800-328-2560
Website: www.agsnet.com

Purpose of Measure

A summary of the K-ABC is also included within the Cognitive Assessment section of this compendium. Some of the information in the two summaries is the same, but in this summary we provide more detailed information on the subtest related to mathematics.

As described by instrument publisher:
"The K-ABC is intended for psychological and clinical assessment, psychoeducational evaluation of learning disabled and other exceptional children, educational planning and placement, minority group assessment, preschool assessment, neuropsychological assessment, and research. The battery includes a blend of novel subtests and adaptations of tasks with proven clinical, neuropsychological, or other research-based validity. This English version is to be used with English-speaking, bilingual and nonverbal children" (Kaufman & Kaufman, 1983a, p. 1).

Population Measure Developed With

- The norming sample included more than 2,000 children between the ages of 2 years, 6 months and 12 years, 6 months old in 1981.
- The same norming sample was used for the entire K-ABC battery, including cognitive and achievement components.
- Sampling was done to closely resemble the most recent population reports available from the U.S. Census Bureau, including projections for the 1980 Census results.
- The sample was stratified for each 6-month age group (20 groups total) between the ages of 2 years, 6 months and 12 years, 6 months, and each age group had at least 100 subjects.
- These individual age groups were stratified by gender, geographic region, SES (as gauged by education level of parent), race/ethnicity (white, black, Hispanic, other), community size, and educational placement of the child.

- Educational placement of the child included those who were classified as speech-impaired, learning-disabled, mentally retarded, emotionally disturbed, other, and gifted and talented. The sample proportions for these closely approximated national norms, except for speech-impaired and learning-disabled children, who were slightly under-represented compared to the proportion within the national population.

Age Range Intended For
Ages 2 years, 6 months through 12 years, 6 months. The Arithmetic subtest can be administered to children ages 3 years and higher.

Key Constructs of Measure
There are two components of the K-ABC (the Mental Processing Scales and the Achievement Scale) and a total of 16 subtests. The assessment yields four Global Scales:

- *Sequential Processing Scale*: Entails solving problems where the emphasis is on the order of stimuli.
- *Simultaneous Processing Scale*: Requires using a holistic approach to integrate many stimuli to solve problems.
- *Mental Processing Composite Scale*: Combines the Sequential and Simultaneous Processing Scales, yielding an estimate of overall intellectual functioning.
- *Achievement Scale*: Assesses knowledge of facts, language concepts, and school-related skills such as reading and arithmetic.

In this summary, we focus on the Arithmetic subtest (from the Achievement Scale), which is administered to children who are 3 years or older.

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
K-ABC utilizes basals and ceilings. The child's chronological age is used to determine the starting item in each subtest. To continue, the child must pass at least one item in the first unit of items (units contain two or three items). If the child fails all items in the first unit, the examiner then starts with the first item in the subtest (unless he/she started with the first item—in that case, the subtest is stopped). In addition, there is a designated stopping point based on age. However, if the child passes all the items in the last unit intended for the child's chronological age, additional items are administered until the child misses one item.

The child responds to requests made by the examiner. The child is required to give a verbal response, point to a picture, build something, etc. For the Arithmetic subtest, the child is asked

to demonstrate knowledge of numbers and mathematical concepts, counting and computation, and other arithmetic abilities.

Who Administers Measure/Training Required?

Test Administration:

- "Administration of the K-ABC requires a competent, trained examiner, well versed in psychology and individual intellectual assessment, who has studied carefully both the *K-ABC Interpretive Manual* and [the] *K-ABC Administration and Scoring Manual*. Since state requirements vary regarding the administration of intelligence tests, as do regulations within different school systems and clinics, it is not possible to indicate categorically who may or may not give the K-ABC" (Kaufman & Kaufman, 1983a, p. 4).

  "In general, however, certain guidelines can be stated. Examiners who are legally and professionally deemed competent to administer existing individual tests…are qualified to give the K-ABC; those who are not permitted to administer existing intelligence scales do not ordinarily possess the skills to be K-ABC examiners. A K-ABC examiner is expected to have a good understanding of theory and research in areas such as child development, tests and measurements, cognitive psychology, educational psychology, and neuropsychology, as well as supervised experience in clinical observation of behavior and formal graduate-level training in individual intellectual assessment" (Kaufman & Kaufman, 1983a, p. 4).

Data Interpretation:

- (Same as above.)

Setting (e.g., one-on-one, group, etc.)

One-on-one.

Time Needed and Cost

Time:

- The time it takes to administer K-ABC increases with age because not all of the subtests are administered at each age. The administration time for the entire battery increases from about 35 minutes at age 2 years, 6 months to 75-85 minutes at ages 7 and above. (The manuals do not provide time estimates for subtests or scales.)

Cost:

- Complete kit: $433.95
- Two Manual set (*Administration and Scoring Manual* and *Interpretive Manual*): $75.95

III. Functioning of Measure

Reliability Information from Manual

Split-Half Reliability

Because of the basal and ceiling method used in the K-ABC, split-half reliability was calculated by taking the actual test items administered to each subject and dividing them into comparable halves, with odd number questions on one half and even numbers on the other. Scale scores were calculated for each half and correlated with each other, and a correction formula was applied in order to estimate reliabilities for full-length tests. Split-half reliabilities for the Arithmetic subtest were .85 at age 3, .89 at age 4, and .89 at age 5 (See Kaufman & Kaufman, 1983b, p. 82).

Test-Retest Reliability

The K-ABC was administered twice to 246 children, two to four weeks after the first administration. The children were divided into three age groups (2 years, 6 months through 4; 5 through 8; and 9 through 12 years, 6 months). For the youngest group, the test-retest correlation of scores on the Arithmetic subtest was .87, (See Kaufman & Kaufman, 1983b, p. 87).

Validity Information from Manual

Construct Validity

Raw scores on all of the K-ABC subtests, as well as the Global Scales, increase steadily with age. Kaufman and Kaufman (1983b, p. 100) describe such a pattern of age-related increases as necessary, but not sufficient, to support the construct validity of any test purporting to be a measure of achievement or intelligence.

The authors also examined internal consistency of the Global Scales as an indicator of construct validity. Each of the subtests was correlated with Global Scale total scores for the entire standardization sample. At age 3, the correlation between the Arithmetic subtest and the Achievement Global Scale was .70; at age 4, it was .77; and at age 5, it was .83. (See Kaufman & Kaufman, 1983b, p. 104).

Concurrent and Predictive Validity

A number of studies were reported by Kaufman and Kaufman (1983b) investigating associations between scores on the K-ABC and scores on other measures of cognitive functioning, achievement, or intelligence. Several of these studies, using various types of samples, were conducted to investigate correlations between the K-ABC scales and Stanford-Binet scores. However, Kaufman and Kaufman do not present correlations for the Arithmetic subtest alone, either with IQ or with the Quantitative subscale of the SB-IV; instead, correlations for Achievement Scale standard scores and other Global Scale standard scores with SB-IV IQ scores are provided. These associations are thus more relevant to the validity of the K-ABC as a general cognitive and achievement measure and are summarized in the K-ABC Cognitive profile.

Reliability/Validity Information from Other Studies

Quite a few studies have looked at the psychometric properties of the K-ABC scale scores, although we found none that looked at the Arithmetic subtest in particular. Thus, these studies are more relevant to the validity of the K-ABC as a general cognitive and achievement measure and are summarized in the K-ABC Cognitive profile.

Comments

- Information presented by Kaufman and Kaufman (1983a) on the Arithmetic subtest indicate strong internal consistency reliability of this subtest. Further, high test-retest correlations indicate a high level of consistency in children's relative performance on repeated administrations of this test across a short time interval.
- With respect to construct validity, high correlations were found between the Arithmetic subtest and Achievement , suggesting that achievement within the math domain (as tapped by the Arithmetic subtest) is strongly associated with other achievement areas assessed with the K-ABC.
- High correlations between the Arithmetic subtest and Achievement Global Scale scores suggest that achievement within the math domain (as tapped by the Arithmetic subtest) is strongly associated with other achievement areas assessed with the K-ABC, providing some support for the construct validity of the Arithmetic subtest as a measure of achievement, although there is less evidence provided for the construct validity of the Arithmetic subtest as an assessment of a unique achievement domain. We found no reports of the K-ABC Arithmetic subtest being used specifically as a measure of achievement or ability within the math domain. The usefulness of this test for this purpose should be explored further.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

None found.

V. Adaptations of Measure

None found.

Early Childhood Measures: Math

Early Childhood Measures: Math

Peabody Individual Achievement Test—Revised (PIAT-R), Mathematics Subtest

I. Background Information

Author/Source

Source: Markwardt, F. C. (1998). *Peabody Individual Achievement Test—Revised: Manual* (Normative Update). Circle Pines, MN: American Guidance Service.

Publisher: American Guidance Service, Inc.
4201 Woodland Road
Circle Pines, MN 55014-1796
Phone: 800-328-2560
Website: www.agsnet.com

Purpose of Measure

A summary of the PIAT-R is also included within the Cognitive Assessment section of this compendium. Some of the information in the two summaries is the same, but in this summary, we pay particular attention to the subtest related to mathematics.

As described by instrument publisher:
"PIAT-R scores are useful whenever a survey of a person's scholastic attainment is needed. When more intensive assessment is required, PIAT-R results assist the examiner in selecting a diagnostic instrument appropriate to the achievement level of the subject. The PIAT-R will serve in a broad range of settings, wherever greater understanding of an individual's achievement is needed. Teachers, counselors, and psychologists, working in schools, clinics, private practices, social service agencies, and the court system, will find it helpful" (Markwardt, 1998, p. 3).

According to the publisher, the uses of PIAT-R include individual evaluation, program planning, guidance and counseling, admissions and transfers, grouping students, follow-up evaluation, personnel selection and training, longitudinal studies, demographic studies, basic research studies, program evaluation studies, and validation studies.

Population Measure Developed With

- The PIAT-R was standardized to be representative of students in the mainstream of education in the United States, from kindergarten through Grade 12.
- A representative sample of 1,563 students in kindergarten through Grade 12 from 33 communities nationwide was tested. The sample included 143 kindergartners. The initial testing was done in the spring of 1986. An additional 175 kindergarten students were tested at 13 sites in the fall of that year to provide data for the beginning of kindergarten.
- Ninety-one percent of the students were selected from public schools, and special education classes were excluded.
- The standardization was planned to have equal numbers of males and females and to have the same proportional distribution as the U.S. population on geographic region, socioeconomic status, and race/ethnicity.

Age Range Intended For
Kindergarten to high school (ages 5years through 18 years). Only the appropriate subsets are administered to each specific age group.

Key Constructs of Measure
The PIAT-R consists of six content areas subtests:

- *Mathematics*: The focus of this summary, this subtest measures students' knowledge and application of mathematical concepts and facts, ranging from recognizing numbers to solving geometry and trigonometry problems.
- *General Information*: Measures students' general knowledge.
- *Reading recognition*: An oral test of reading that measures children's ability to recognize the sounds associated with printed letters and their ability to read words aloud.
- *Reading Comprehension*: Measures students' understanding of what is read.
- *Spelling*: Measures students' ability to recognize letters from their names or sounds and to recognize standard spellings by choosing the correct spelling of a word spoken by the examiner.
- *Written Expression*: Assesses children's written language skills at two levels. Level 1 is appropriate for kindergarten and first-grade subjects, and Level 2 is appropriate for Grades 2 through 12. Level 1 tests pre-writing skills such as copying and writing letters, words, and sentences from dictation.

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

Comments

- The following limitations of the PIAT-R are cited in the manual:
  - The test is not designed to be used as a diagnostic test.
  - The test identifies a person's general level of achievement but is not designed to provide a highly precise assessment of achievement.
  - The items in the test present a cross section of curricula used across the United States and are not designed to test the curricula of a specific school system.
  - Administration and interpretation of the test scores require different skills. The manual cautions that only those with appropriate skills should engage in interpretation of scores.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
As the PIAT-R is administered to such a wide age range of respondents and contains a range of questions that vary greatly in difficulty, the examiner must determine a *critical range*. The

critical range includes those items of appropriate difficulty for the individual's level of achievement. Details on how to determine the *critical range* are provided in the PIAT-R manual. PIAT-R utilizes basals and ceilings.

The Mathematics subtest uses a multiple-choice format. It consists of 100 questions ranging in difficulty from "discriminating and matching tasks" to "geometry and trigonometry content." For the first 50 items, the examiner reads the question while the response choices are displayed to the student. The student may respond either by pointing or saying the quadrant number of the correct answer. For the last 50 items, the questions are shown as well as read to the student. The examiner records and immediately scores the child's oral response to each item.

Who Administers Measure/ Training Required?

Test Administration:

- Any individual who learns and practices the procedures in the PIAT-R manual can become proficient in administering the test. Each examiner should study Part II and Appendix A of the manual, the test plates, the test record, and the Written Expression Response Booklet.

Data Interpretation:

- Individuals with knowledge and experience in psychology and education, such as psychologists, teachers, learning specialists, counselors, and social workers are the most appropriate candidates for interpreting scores. Interpretation requires an understanding of psychometrics, curriculum, and the implications of a subject's performance.

Setting (e.g. one-on-one, group, etc)

One-on-one.

Time Needed and Cost

Time:

- There is no time limit on the test (except for Level II of the Written Expression subtest), and the manual does not provide an estimate for the Mathematics subtest on its own. Typically all six subtests can be administered in one hour. Items are scored while the subtests are being administered (excluding Written Expression).

Cost:

- Complete kit: $342.95
- Manual: $99.95

Comments

- The PIAT-R is designed to be administered with all six subtests in a specific order. All six subtests should be administered in order to ensure maximum applicability of the norms. Separate administration of the Mathematics subtest does not follow this recommendation.
- If the student to whom the test is being administered is young, it may be necessary to do Training Exercises (provided at the beginning of each subtest) to instruct the child on how to point as the appropriate method of responding to the multiple choice questions.

III. Functioning of Measure

Reliability Information from Manual

Split-Half Reliability

For each subtest, estimates were obtained by correlating the total raw score on the odd items with the total raw score on the even items. Correlations were corrected using the Spearman-Brown formula to estimate the reliabilities of full-length tests. The manual presents results both by grade level and by age. For the kindergarten subsample, the Mathematics subtest reliability was .84 (see Markwardt, 1998, p.59).

Test-Retest Reliability

Students were randomly selected from the standardization sample. Fifty subjects were selected in each of grades kindergarten, 2, 4, 6, 8, and 10. Participants were retested from 2 to 4 weeks after the initial assessment. For the kindergarten subsample, the test-retest reliability estimate for the Mathematics subtest was .89 (see Markwardt, 1998, p.61).

Other Reliability Analyses

A total of four different reliability analyses were reported. In addition to split-half and test-retest reliabilities (summarized above), Kuder-Richardson and item response theory methods were used to estimate reliability. Results of these analyses (conducted both by grade and by age) parallel the split-half and test-retest reliability results (see Markwardt, 1998, pp. 59-63).

Validity Information from Manual

Construct Validity

- According to Markwardt (1998, p. 66), "The extent to which test scores show a progressive increase with age or grade is a major criterion for establishing the validity of various types of ability and achievement tests." In the standardization sample, mean and median scores on the Mathematics subtest of the PIAT-R demonstrated age- and grade-related increases through age 17 and grade 11 (pp. 54-55).
- No other information was provided regarding the validity of the Mathematics subtest No studies were reported in which Mathematics subtest scores were associated with other measures of mathematical achievement or ability that could provide support for the concurrent or predictive validity of the subtest. Correlations between scores on PIAT-R Mathematics subtest and on the Peabody Picture Vocabulary Test—Revised (PPVT-R) were

reported for a sample including 44 5-year-olds and 150 6-year-olds. These correlations were .51 at age 5 and .55 at age 6 (see Markwardt, 1998, p.66).

Reliability/Validity Information from Other Studies

None found.

Comments

- Information provided by Markwardt (1998) indicates high internal consistency (split-half reliability) and test-retest reliability.
- Intercorrelations of the different subtests indicate that mathematics achievement as tapped by the PIAT-R was moderately associated with achievement in other areas (i.e. reading, spelling, and general information), but that a substantial amount of unique information may be obtained from this subtest as well. Limited information is presented about the validity of the Mathematics subtest. As noted above, validity data are not provided at the level of the subtest. Further, as noted in the PIAT-R profile within the Cognitive Assessment section of this compendium, in general the test developers present limited validity information. In particular, concurrent validity for the PIAT-R is reported only in relation to the PPVT-R, and there is no examination of validity with respect to aspects of cognitive development or achievement other than language (although an appendix in the manual summarizes studies that have been conducted using the original PIAT).

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

Blau (1999) conducted a study of the quality of child care using data from the National Longitudinal Survey of Youth (NLSY). The original sample consisted of 12,652 youth who were 14- to 21-years-old in 1979. Beginning in 1986, the children of female sample members were assessed yearly between the ages of 4 and 11. The assessments included the Mathematics subtest of the original PIAT. Measures of the quality of child care were mothers' reports of group size, staff-child ratio, and caregiver training. Blau found that group size and staff-child ratio were uncorrelated with PIAT outcomes, but training was positively and significantly correlated. However, after further variables were taken into account (i.e., number of arrangements, type of care, hours per week), these associations were no longer significant .

V. Adaptations of Measure

None found.

Early Childhood Measures: Math

Early Childhood Measures: Math

Stanford-Binet Intelligence Scale, Fourth Edition[9], Quantitative subtest

I. Background Information

Author/Source

Source: Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986a). *Stanford-Binet Intelligence Scale: Fourth Edition. Guide for administering and scoring*. Itasca, IL: The Riverside Publishing Company. (See also Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986b). *Stanford-Binet Intelligence Scale: Fourth Edition. Technical manual*. Itasca, IL: The Riverside Publishing Company.)

Publisher: Riverside Publishing
425 Spring Lake Drive
Itasca, IL 60143-2079
Phone: 800-323-9540
Website: www.riverpub.com

Purpose of Measure

This profile focuses on the Stanford-Binet Intelligence Scale (Fourth Edition), Quantitative subtest. A profile of the scale as a whole is included within Cognitive Assessment section of this compendium.

Purpose of the Stanford-Binet Intelligence Scale (Fourth Edition) as a whole, as described by instrument publisher:

"The authors have constructed the Fourth Edition to serve the following purposes:

1. To help differentiate between students who are mentally retarded and those who have specific learning disabilities.
2. To help educators and psychologists understand why a particular student is having difficulty learning in school.
3. To help identify gifted students.
4. To study the development of cognitive skills of individuals from ages 2 to adult" (Thorndike, Hagen, & Sattler, 1986a, p. 2).

Population Measure Developed With

- One sample was used to standardize all of the subtests.
- The sampling design for the standardization sample was based on five variables, corresponding to 1980 Census data. The variables were geographic region, community size, ethnic group, age, and gender.
- Information on parental occupation and educational status was also obtained.
- The sample included 5,013 participants from ages 2 to 24. Included in this sample were 226 2-year-olds; 278 3-year-olds; 397 4-year-olds; and 460 5-year-olds.

[9] Stanford-Binet Intelligence Scale, Fifth Edition will be released in 2003.

Age Range Intended For
Ages 2 years through adulthood.

Key Constructs of Measure
The SB - IV contains 15 subtests covering four areas of cognitive ability:

- *Verbal Reasoning*: Vocabulary, Comprehension, Absurdities, Verbal Relations.
- *Quantitative Reasoning*: Quantitative, Number Series, Equation Building.
- *Abstract/Visual Reasoning*: Pattern Analysis, Copying, Matrices, Paper Folding and Cutting.
- *Short-term* Memory: Bead Memory, Memory for Sentences, Memory for Digits, Memory for Objects.

Subtests can be administered individually or in various combinations to yield composite Area Scores and a total Composite score for the test. For this profile, we will focus on the Quantitative subtest, the only Quantitative Reasoning Area subtest that can be administered to 2- to 5-year-old children (Number Series can generally be administered starting at age 7; Equation Building is generally administered at ages 12 and higher). Raw scores for subtests and Areas (including Quantitative) are converted to Standard Age Scores in order to make scores comparable across ages and across different tests.

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

Comments

- While the quantitative aspect of cognitive development is addressed through multiple subtests at older ages, for very young children it is based on a single subtest.

II. Administration of Measure

Who is the Respondent to the Measure?
Individuals aged 2 years through adulthood.

If Child is Respondent, What is Child Asked to Do?
SB- IV utilizes basals and ceilings within each subtest, based on sets of four items. A basal is established when the examinee passes all of the items in two consecutive sets. A ceiling is established when the examinee fails at least three out of four items in two consecutive sets.

Guidelines for the tests to be administered are not provided based on age, but on the entry level of the examinee. Entry level is determined through a combination of the score on the Vocabulary subtest and chronological age. We focus on one math subtest here—Quantitative. However, neither the *Technical Manual* nor the *Guide for Administering and Scoring the Fourth Edition* provide examples of the items included in this subtest.

Who Administers Measure/Training Required?

Test Administration:

- "Administering the Stanford-Binet scale requires that you be familiar with the instrument and sensitive to the needs of the examinee. Three conditions are essential to securing accurate test results: (1) following standard procedures, (2) establishing adequate rapport between the examiner and the examinee, and (3) correctly scoring the examinee's responses" (Thorndike, Hagen, & Sattler, 1986a, p. 9).
- The manual does not provide guidelines for examiners' education and experience.

Data Interpretation:

- The manual does not specify the education and experience need for data interpretation using the SB-IV.

Setting (e.g., one-on-one, group, etc.)

This test is designed to be administered in a one-on-one setting.

Time Needed and Cost

Time:

- Time limits are not used. "Examinees vary so markedly in their test reactions that it is impossible to predict time requirements" (Thorndike, Hagen, & Sattler, 1986a, p. 22).

Cost:

- Examiner's Kit: $777.50
- *Guide for Administering and Scoring Manual*: $72.50
- *Technical Manual*: $33

Comments

- The SB-IV utilizes an adaptive-testing format. Examinees are administered a range of tasks suited to their ability levels. Ability level is determined from the score on the Vocabulary subtest, along with chronological age.

- At ages 4 and above, the range of item difficulty is large, so either a zero score or a perfect score on any subtest is very infrequent. However, at age 2, zero scores occur frequently on certain subtests due to an inability to perform the task or a refusal to cooperate. According to the manual, the SB-IV does not discriminate adequately among the lowest 10 to 15 percent of the 2-year-old group. At age 3, SB - IV adequately discriminates among all except the lowest two percent.

III. Functioning of Measure

Reliability Information from Manual

Internal Consistency

Split-half reliabilities of the subtests were calculated using the Kuder-Richardson Formula 20 (KR-20). All items below the basal level were assumed to be passed, and all items above the

ceiling level were assumed to be failed. The manual provides reliability data for every age group, but we focus on the data for ages 2 years to 5 years. For the Quantitative subtest, at age 2, the split-half reliability estimate was .81; at age 3, it was .84; and at age 5, it was .88 (Thorndike, Hagen, & Sattler, 1986b, p. 40).

Test-Retest Reliability
Test-retest reliability data were obtained by retesting a total of 112 children, 57 of whom were first tested at age 5. The length of time between administrations varied from 2 to 8 months, with an average interval of 16 weeks. The age 5 subsample consisted of 29 boys and 28 girls; 65 percent were white, 31 percent were black, 2 percent were Hispanic, and 2 percent were Native American. For the age-5 subsample, the test-retest reliability of the Quantitative subtest was .71 (Thorndike, Hagen, & Sattler, 1986b, p. 46).

Validity Information from Manual
Construct Validity
Correlations were calculated between all subtest, area, and composite scores (see Thorndike, Hagen, & Sattler, 1986, p. 110-113). Because the Quantitative subtest is the only math-related subtest that can be administered to preschoolers, it is not possible to determine if correlations between the Quantitative subtest and other Quantitative Reasoning Area subtests might have been higher than the correlations between the Quantitative subtest and subtests from other areas (i.e., Verbal Reasoning Area subtests, Abstract/Visual Reasoning Area subtests, and Short-Term Memory Area subtests). Correlations between Quantitative subtest scores and Area scores ranged from .29 (Short-Term Memory) to .45 (Verbal Reasoning) at age 2; from .50 (Verbal Reasoning) to .59 (Abstract/Visual Reasoning) at age 3; from .63 (Verbal Reasoning and Short-Term Memory) to .67 (Abstract/Visual Reasoning) at age 4; and from .63 (Verbal Reasoning) to .68 (Abstract/Visual Reasoning) at age 5. Correlations between Quantitative subtest scores and Composite scores were .72 at age 2, .80 at age 3, .87 at age 4, and .86 at age 5. It is interesting to note that no other single subtest correlated more highly with the Composite than did the Quantitative subtest at any age (although two other subtests—Comprehension and Pattern Analysis—also correlated .72 with the Composite at age 2).

Concurrent Validity
Several studies were conducted comparing SB-IV scores to scores on other assessments. We focus here on the study with the youngest sample, in which Standard Age Scores on the SB-IV were correlated with Verbal, Performance, and Full Scale IQ scores derived from the Wechsler Preschool and Primary Scale of Intelligence (WPPSI). The sample consisted of 75 participants with a mean age of 5 years and 6 months. Thirty-four children were male, 41 were female. Eighty percent of the sample was white, 7 percent was black, 7 percent were Asian, and the remainder were classified as other race/ethnicity. Thorndike, Hagen, & Sattler (1986) expected that SB-IV Quantitative Reasoning would be more highly associated with WPPSI Verbal IQ than with Performance IQ. Findings supported this expectation to some extent, although the difference in correlations was very small (.70 vs. .66; see Thorndike, Hagen, & Sattler, 1986, p. 64). Quantitative scores also correlated .73 with WPPSI Full Scale IQ scores.

Reliability/Validity Information from Other Studies

We found few studies that examined the psychometric properties of the Quantitative subtest alone. The following studies examined the characteristics of the full battery (see the SB-IV Cognitive profile for further studies examining the reliability or validity of the full battery)

- In one study relevant to the Quantitative subtest, Johnson, Howie, Owen, Baldwin, and Luttman (1993) investigated the usefulness of the SB-IV with young children. The sample consisted of 121 3-year-olds; 52 girls and 69 boys. The sample included both white and black children (proportions not given). The eight SB-IV subtests appropriate for 3-year-olds were administered. The investigators found that 55 percent of the children were unable to obtain a score (that is, they did not get a single item correct) on some SB-IV subtests. One of the most problematic subtests for obtaining a score was the Quantitative subtest, which should be a cause for concern when using this subtest with young children. However, it is not clear whether this pattern of findings was specific to the particular sample and administration of the measure, or may be a more general problem with the measure.
- Krohn and Lamp (1989) studied the concurrent validity of the Kaufman Assessment Battery for Children (K-ABC) and the SB-IV, both compared to a previous version of the Stanford-Binet Intelligence Scale, Form LM (SB-LM; the third edition of the assessment). The sample consisted of 89 Head Start children, ranging in age from 4 years, 3 months to 6 years, 7 months, with a mean age of 4 years, 1 month. Fifty children were white and 39 were black. The authors found that K-ABC and SB-IV scores were significantly associated with scores on the to SB-LM, supporting the concurrent validity of both the SB-IV and the K-ABC.
- Gridley and McIntosh (1991) explored the underlying factor structure of SB-IV. The study utilized two samples—50 2- to 6-year-olds, and 137 7- to 11-year-olds. Altogether, 90 percent of the subjects where white, and 10 percent were black. The eight subtests appropriate for use with younger ages were administered to the younger sample. Among 2- to 6-year-olds, the authors found more support for a two-factor model (Verbal Comprehension and Nonverbal Reasoning/Visualization) or three-factor model (Verbal Comprehension, Nonverbal Reasoning/Visualization, and Quantitative) than for the four-factor model posited to exist by the test developers (i.e., Verbal Reasoning, Abstract/Visual Reasoning, Quantitative Reasoning, and Short-Term Memory), thus providing a limited degree of support for Quantitative Reasoning as a separate and distinct area of ability tapped by the SB-IV.

Comments

- Information provided by Thorndike, Hagen, & Sattler (1986b) indicates strong internal consistency of the Quantitative subtest of the SB-IV at ages 2 through 5, although there appears to be a slight trend for internal consistency to increase somewhat across this age period. Test-retest reliability was not assessed for the youngest ages. At age 5 the strong correlation across testing session suggesting a high level of consistency in children's relative scores on the Quantitative subtest across an average time span of approximately four months.
- Because there is a single math-related subtest administered to preschoolers, it is difficult to determine whether correlations between the Quantitative subtest and Area scores support the construct validity of the subtest as a measure of math-related ability. Strong correlations with the Composite, as well as the increasing strength of correlations between the

Quantitative subtest and the Area and Composite scores between ages 2 and 4 may support the validity of the Quantitative subtest as a key component of general cognitive ability as assessed with the SB-IV.

- With respect to results presented related to concurrent validity, the extent to which results presented by Thorndike, Hagen, & Sattler (1986b) can be used as evidence of the validity of the Quantitative Reasoning subtest as a measure of mathematical/quantitative ability is limited, given that the WPPSI measures were not specifically related to math or quantitative reasoning.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- We found no studies that utilize the Quantitative subtest alone, or that described results pertaining specifically to the Quantitative subtest (see the SB-IV Cognitive profile for reports of studies employing the SB-IV and the SB-LM for evaluations of intervention programs).

V. Adaptations of Measure

None found.

Early Childhood Measures: Math

Early Childhood Measures: Math

Test of Early Mathematics Ability—Second Edition (TEMA-2)

I. Background Information

Author/Source

Source: Ginsburg, H. P., & Baroody, A. J. (1990). *Test of Early Mathematics Ability, Second Edition: Examiner's manual*. Austin, TX: PRO-ED, Inc.

Publisher: PRO-ED, Inc.
8700 Shoal Creek Blvd.
Austin, TX 78757-6897
Phone: 800-897-3202
Website: www.proedinc.com

Purpose of Measure

As described by instrument publisher:
The TEMA serves several purposes: "1. Identify those children who are significantly behind or ahead of their peers in the development of mathematical thinking; 2. identify specific strengths and weaknesses in mathematical thinking; 3. suggest instructional practices appropriate for individual children; 4. document children's progress in learning arithmetic; and 5. serve as a measure in research projects" (Ginsburg & Baroody, 1990, p. 4).

Population Measure Developed With

- The normative sample for TEMA-2 consisted of 896 children in 27 states representing all regions of the United States.
- Children in the sample ranged in age from 3 to 8years.
- The sample was located in three ways. First, the test developers found a nationwide group of professionals who had purchased tests from PRO-ED. They were asked to test 20 to 30 children in their areas using TEMA-2. Second, individuals across the country who had assisted in the development of other PRO-ED tests were asked to test 20 to 30 children. Third, teams of examiners were trained by the authors to collect data from sites in the four major census districts.
- The normative sample was representative of the national population in regard to sex, race (white, black, and other), geographic region of residence, residence in an urban or rural community, and parent occupation (white-collar, blue-collar, service, farm, or other).

Age Range Intended For

Ages 3 years through 8 years, 11 months.

Key Constructs of Measure

The TEMA-2 measures both formal mathematics (skills and concepts learned in school) and informal mathematics (concepts learned outside of school). Formal math constructs include conventions, number facts, concepts, and calculations. Informal math constructs include relative magnitude, counting, and calculation.

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

Comments

- No information on family income or children's grade in school was presented for the norming sample.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
Basals and ceilings are used in TEMA-2. Testing begins with an item corresponding to the child's age. Testing continues until five consecutive items are missed or until the last item is administered. If the child does not answer five consecutive items correctly, the examiner returns to the entry point and tests downward until five items in a row are answered correctly or until the first item is administered. All items below the basal are scored as correct.

The child is asked to count objects in a picture, show a certain number of fingers, indicate by pointing which of two cards has more objects than the other, write numbers, perform mental addition/subtraction, and other mathematics-related activities

Who Administers Measure/Training Required?
Test Administration:

- Examiners who administer TEMA-2 should have formal training in assessment, such as college coursework or assessment workshops.

Data Interpretation:

- Examiners who interpret TEMA-2 results should have formal training in statistics and in the procedures governing test administration, scoring, and interpretation.

Setting (e.g., one-on-one, group, etc.)
One-on-one.

Time Needed and Cost
Time:

- Administration takes about 20 minutes.

Cost:

- Complete kit: $169
- Examiner's Manual: $46

Comments

- TEMA-2 may be given in more than one sitting if needed.

III. Functioning of Measure

Reliability Information from Manual

Internal Reliability

Split-half reliabilities were estimated by calculating coefficient alphas separately for each one-year age group within the standardization sample (ages 3 through 8). The average coefficient alpha across the six age groups was .94. Alphas for separate age groups were highly consistent, ranging from .92 to .96 (see Ginsburg, & Baroody, 1990, p. 34).

Test-Retest Reliability

Test-retest reliability of the original TEMA was examined by assessing 71 4- and 5-year-olds in preschools and day care centers in Austin, Texas. The TEMA was administered twice, with one week between test administrations. The partial correlations between scores for the two test administrations, controlling for age, was .94 (see Ginsburg, & Baroody, 1990, p. 34).

Validity Information from Manual

With the exception of information on age-related changes in TEMA-2 raw scores, none of the validity information provided by Ginsburg and Baroody (1990) actually utilized the TEMA-2. Results from studies utilizing the original TEMA were reported, as were results from studies using an abbreviated version of the TEMA-2, the Math subtest of the Screening Children for Related Early Educational Needs (SCREEN) assessment (Hresko, Reid, Hammill, Ginsburg, & Baroody, 1988).

Concurrent Criterion-Related Validity

Ginsburg and Baroody (1990) report two studies in which scores on the TEMA or TEMA-2 short form were correlated with scores on other measures of math abilities. In one study, standard scores on the TEMA were correlated with standard scores on the Math Calculation subtest from the Diagnostic Achievement Battery (Newcomer & Curtis, 1984) in a sample of 23 6-year-olds and 17 8-year-olds from one elementary school. Correlations, corrected for attenuation, were .40 for the younger children and .59 for the older children (p. 35). According to Ginsburg and Baroody, "…one might conclude that the findings support the criterion-related validity of the test" (p. 35).

In a second study, the short form of the TEMA-2 was administered to 35 6-year-old children, along with the Math subtest of the Quick Score Achievement Test (Q-SAT; Hammill, Ammer, Cronin, Mandelbaum, & Quinby, 1987). The correlation between these two measures, .46, was very similar to that reported in the previous study (see Ginsburg & Baroody, 1990, p. 25).

Construct Validity

Ginsburg and Baroody (1990) briefly present different types of evidence in support of the construct validity of the TEMA-2, including age differentiation, significant associations with tests of school achievement, and significant associations with aptitude tests. With regard to age differentiation, the authors suggest that because the TEMA-2 is designed to measure math-

related abilities that increase with age, raw scores should increase with age. This pattern of scores was in fact reported, with mean raw scores steadily increasing from 5.24 at age 3 to 46.32 at age 8. Further, TEMA-2 raw scores were found to correlate .83 with age (see Ginsburg & Baroody, 1990, p. 36).

The second type of evidence for construct validity presented by Ginsburg and Barrody (1990) involved relationships of the TEMA-2 to other measures of school achievement, based on the view that measures of achievement should be significantly associated with each other even when the specific areas of achievement tapped by the measures differ. TEMA-2 scores were correlated with scores on the Test of Early Language Development (TELD; Hresko, Reid, & Hammill, 1981) in a sample of 62 4- and 5-year-olds in day care centers in Austin, TX. The correlation between these two measures, controlling for child age, was .39. In a separate study, TEMA-2 short form scores were correlated with scores on other subtests of the SCREEN. Correlations between these subtest scores were .95 with Language, .96 with Reading, and .87 with Writing. These correlations were interpreted by Ginsburg and Baroody (1990) as providing "…solid evidence of the TEMA-2 score's construct validity" (p. 36).

Ginsburg and Baroody (1990) also indicated that significant correlations between TEMA-2 and measures of academic aptitude would support the construct validity of the TEMA-2. The relationship between TEMA scores and scores on the Slosson Intelligence Test (SIT, second edition; Slosson, 1983) was examined in a sample of 62 4- and 5-year-olds (no other sample details are given). The correlation between math ability as tapped by the TEMA (and TEMA-2) and intelligence as tapped by the SIT was .66 (see Ginsburg & Baroody, 1990, p. 36).

Reliability/Validity Information from Other Studies
None found.

Comments

- Predictive validity (for example, relating the TEMA-2 to subsequent school achievement in mathematics or later scores on mathematics assessments) is not reported on in manual.
- As noted earlier, much of the research relevant to the validity of the TEMA-2 has actually been conducted with the original TEMA or with an abbreviated version of the TEMA-2 that is incorporated into another measure.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

None found.

V. Adaptations of Measure

A short form of the TEMA-2 that was included as part of the Screening Children for Related Early Educational Needs (SCREEN) assessment (Hresko, Reid, Hammill, Ginsburg, & Baroody,

1988) was described in the manual. Some of the validity information provided for the TEMA-2 actually involved the use of this abbreviated measure (see above).

Early Childhood Measures: Math

Early Childhood Workshop: Math

Woodcock-Johnson III (WJ III), Tests of Achievement

I. Background Information

Author/Source

Source: McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III: Technical manual*. Itasca, IL: The Riverside Publishing Company. (See also Mather, N., & Woodcock, R. W. (2001a). *Woodcock-Johnson III Tests of Achievement: Examiner's manual*. Itasca, IL: The Riverside Publishing Company; Mather, N., & Woodcock, R. W. (2001b). *Woodcock-Johnson III Tests of Cognitive Abilities: Examiner's manual*. Itasca, IL: The Riverside Publishing Company.)

Publisher: Riverside Publishing
425 Spring Lake Drive
Itasca, IL 60143-2079
Phone: 800-323-9540
Website: www.riverpub.com

Purpose of Measure

A summary of WJ III is also included within the Cognitive Assessment section of this compendium. Some of the information in the two summaries is the same, but in this summary we provide more detailed information on the subtests related to mathematics.

As described by instrument publisher:
The purpose of the WJ III is to determine the status of an individual's academic strengths and weaknesses. WJ III Tests of Achievement can serve as an in-depth evaluation after an individual has failed a screening assessment. They can also be used to make decisions regarding educational programming for individual children. The authors also suggest that they can be used for program evaluation and research.

"The WJ III batteries were designed to provide the most valid methods for determining patterns of strengths and weaknesses based on actual discrepancy norms. Discrepancy norms can be derived only from co-normed data using the same participants in the norming sample. Because all of the WJ III tests are co-normed, comparisons among and between a participant's general intellectual ability, specific cognitive abilities, oral language, and achievement scores can be made with greater accuracy and validity than would be possible by comparing scores from separately normed instruments" (McGrew & Woodcock, 2001, p. 4).

Population Measure Developed With

- The norming sample for WJ III consisted of a nationally representative sample of 8,818 children and adults in 100 U.S. communities. Participants ranged in age from 2 years to 80+ years.
- The preschool sample (ages 2 years to 5 years and not enrolled in kindergarten) had 1,143 children.

- All participants were administered all tests from both the WJ III COG and the WJ III ACH (see description, below).
- Participants were randomly selected within a stratified sampling design taking into account Census region, community size, sex, race and Hispanic origin, type of school, type of college/university, education level, occupational status of adults and occupation of adults in the labor force. Preschool children were selected using a stratified sampling design taking into account region, community size, sex, race and Hispanic origin, as well as parent education and occupation.

Age Range Intended For
Ages 2 years through adulthood (however, some subtests cannot be administered to 2-, 3-, or 4-year-olds).

Key Constructs of Measure
WJ III consists of two batteries—the WJ III Tests of Cognitive Abilities (WJ III COG) and the WJ III Tests of Achievement (WJ III ACH). For this summary, we focus on four subtests of WJ III ACH.

WJ III ACH consists of 22 subtests, four of which are related to mathematics. The tests measure math calculation skills (Test 5, administered to individuals ages 5 and older), math fluency (Test 6; measures the ability to solve simple addition, subtraction, and multiplication problems quickly, administered to individuals ages 7 and older), and math reasoning (Test 10: Applied Problems and Test 18: Quantitative Concepts, both administered at all ages). Tests 5, 6, and 10 are part of the standard battery, while Test 18 is included in the extended battery. Several clusters can be computed—Broad Math (Tests 5, 6, and 10), Math Calculation Skills (Tests 5 and 6) and Math Reasoning (Tests 10 and 18).

Norming of Measure (Criterion or Norm Referenced)
Norm referenced.

Comments
- Not all of the subtests can be administered to 2-, 3-, or 4-year-olds. Therefore, composite scores may not be available for children at each age.

II. Administration of Measure

Who is the Respondent to the Measure?
Individuals aged 2 years through adulthood.

If Child is Respondent, What is Child Asked to Do?
WJ III utilizes basals and ceilings, although the rules are different for each subtest.

- Examples of what the respondent is asked to do:
  - Write a single number.
  - Solve simple arithmetic problems.

- Solve a word problem read aloud to him/her.
- Count and identify numbers, shapes, and sequences.

Who Administers Measure/Training Required?

Test Administration:

- Examiners who administer WJ III should have a thorough understanding of the WJ III administration and scoring procedures. They should also have formal training in assessment, such as college coursework or assessment workshops.

Data Interpretation:

- Interpretation of WJ III requires more knowledge and experience than that required for administering and scoring the test. Examiners who interpret WJ III results should have graduate-level training in statistics and in the procedures governing test administration, scoring, and interpretation.

Setting (e.g., one-on-one, group, etc.)

One-on-one.

Time Needed and Cost

Time:

- The time needed for test administration depends on the number and combination of subtests being administered. Each subtest requires about 5 to 10 minutes.

Cost:

- Complete battery: $966.50
- Achievement battery: $444
- Manual: $52

III. Functioning of Measure

Reliability Information from Manual

Internal Reliability

Internal reliabilities were calculated in one of two ways, depending on the type of subtest. A split-half procedure was used for all math-related tests used for children ages 6 and under.[10] The calculation separated odd and even items, and items below the participant's basal level were scored as correct while items above the ceiling level were scored as incorrect. Scale scores were calculated for each half and correlated with each other, and a correction formula was applied in order to estimate reliabilities for full-length tests. Split-half reliabilities were high for all math tests. For Broad Math, r = .96 at age 5 (this score cannot be calculated for 2-, 3-, or 4-year-olds); for Math Calculation Skills, r = .97 at age 5 (this score cannot be calculated for 2-, 3-, or 4-year-olds); and for Math Reasoning, r = .92 at ages 2 and 3, .94 at age 4, and .95 at age 5 (see McGrew, & Woodcock, 2001, p. 118, 143, 149).

[10] Math Fluency is a timed test requiring a different procedure for estimating internal reliability. Because it is not administered to children below the age of 7, these procedures will not be discussed in this review.

Test-retest reliability
Test-retest reliabilities were reported for the Applied Problems subtest for 1,196 children and adults (total number–ages 2 to 95). Test-retest reliabilities remained high even across extended time intervals. For children ages 2 to 7, the correlation after less than one year was .90. Between one and two years later, the correlation was .85. Between three and ten years later, the correlation was .90 (see McGrew, & Woodcock, 2001, p. 40).

Test-retest reliabilities were also presented for several WJ III ACH subtests and clusters from a second study of 457 children and adults (total number—ages 4 to 95). Participants in this study were re-tested one year after the initial administration. For children ages 4 to 7, the correlation for Calculation was .87 and for Applied Problems, the correlation was .92. The correlation for Math Fluency (administered only to children ages 7 and higher) was .75. For the Broad Math cluster score, the correlation was .94; for the Math Calculation Skills cluster score, it was .89 (see McGrew, & Woodcock, 2001, p. 42-43).

Validity Information from Manual
Internal Validity
Internal structure validity was examined by investigating the patterns of associations among subtests and among cluster scores using confirmatory factor analysis. According to McGrew & Woodcock (2001, p. 59-68 and Appendix F), the expected patterns emerged, with subtests designed to measure similar constructs being more highly associated than were those measuring widely differing constructs. These analyses did not include data from children below the age of 6, however.

Concurrent Validity
A study of 202 young children (mean age of 52.7 months; age range from 1 year, 9 months to 6 years, 3 months) was conducted in South Carolina. Children completed all of the tests from WJ III COG and WJ III ACH that were appropriate for preschoolers. They were also administered the Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R; Wechsler, 1989) and the Differential Abilities Scale (DAS; Elliott, 1990). However, the manual only presents the findings for WJ III COG, not WJ III ACH. Since the math subtests are part of WJ III ACH, it is not possible to report on their concurrent validity.

Reliability/Validity Information from Other Studies
- Very few studies have been published about the psychometrics of WJ III, due to its recent (2001) publication. Many studies have been conducted on the psychometric properties of WJ-R, but we were unable to find any that are relevant to the preschool age range.

Comments
- Reliability and validity information is not provided in the manual for all of the math subtests or for composite scores. Inter-rater reliability was only reported for three of the WJ III subtests overall (all having to do with writing), and thus information on inter-rater reliability for the math subtests is not available. However, the results of studies presented in the WJ III Technical Manual investigating split-half and test-retest reliability of math-related composites indicate that these forms of reliability are strong.

- It is worth noting that the findings presented for test-retest reliabilities used unusually long intervals between tests—from 1 year to 10 years.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- WJ III is a very recent publication of the test, but several studies using the WJ-R have been published. For example, math subtests of the WJ-R have been used in several studies of the quality of child care, including the Cost, Quality, and Outcomes Study (CQO; Peisner-Feinberg & Burchinal, 1997). One hundred, seventy child care centers participated in CQO, and random sampling procedures for children within centers resulted in an analysis sample of 757 children. The mean age was 4 years, 4 months; 15.9 percent were black, 4.6 percent Hispanic, 67.9 percent white, and 11.6 percent other race/ethnicity. The authors used the Applied Problems subtest of the WJ-R to measure children's pre-math skills. They found that children in low quality child care (as measured by a composite index created from four measures) had significantly lower pre-math scores than children in medium- or high-quality care. However, the results did not hold for pre-math scores after family selection factors were controlled for.

V. Adaptations of Measure

Spanish Version of WJ III

A Spanish version of the WJ III is available.

Cognitive (General) and Math References

Bayley, N. (1993). *Bayley Scales of Infant Development, Second Edition: Manual*. San Antonio, TX: The Psychological Corporation.

Blau, D. M. (1999). The effects of child care characteristics on child development. *Journal of Human Resources, 34*(4), 786–822.

Boehm, A.E. (1986a). Boehm Test of Basic Concepts, Revised (Boehm–R). San Antonio, TX: The Psychological Corporation.

Boehm, A.E. (1986b). Boehm Test of Basic Concepts, Preschool version (Boehm–Preschool). San Antonio, TX: The Psychological Corporation.

Boller, K., Sprachman, S., Raikes, H., Cohen, R. C., Salem, M., & van Kammen, W. (2002). *Fielding and analyzing the Bayley II Mental Scale: Lessons from Early Head Start*. Paper prepared for Selecting Measures for Young Children in Large Scale Surveys, a workshop sponsored by the Research Network on Child and Family Well-Being and the Science and Ecology of Early Development, Washington, DC.

Bracken, B. A. (1998). *Bracken Basic Concept Scale—Revised: Examiner's manual*. San Antonio, TX: The Psychological Corporation.

Brooks-Gunn, J., Gross, R.T., Kraemer, H.C., Spiker, D. & Shapiro, S. (1992). Enhancing the cognitive outcomes of low birth weight, premature infants: For whom is the intervention most effective? *Pediatrics, 89*(6), 1209-1215.

Brooks-Gunn, J., Liaw, F. & Klebanov, P.K. (1992). Effects of early intervention on cognitive function of low birth weight preterm infants. *Journal of Pediatrics, 120*, 350-359.

Brunet, O., & Lézine, I. (1951). *Le développement psychologique de la premièere enfance*. Paris: Presses Universitaires.

Burchinal, M.R., Campbell, F.A., Bryant, D.M., Wasik, B.H., & Ramey, C.T. (1997). Early intervention and mediating process in cognitive performance of children of low-income African American families. *Child Development, 68*(5), 935-954.

Burchinal, M. R., Roberts, J.E., Riggins, R., Zeisel, S.A., Neebe, E, & Bryant, D. (2000). Relating quality of center child care to early cognitive and language development longitudinally. *Child Development, 71*(2), 339-357.

Campbell, F. A., Pungello, E. P., Miller-Johnson, S., Burchinal, M., & Ramey, C. T. (2001). The development of cognitive and academic abilities: Growth curves from an early childhood educational experiment. *Developmental Psychology, 37*(2), 231-242.

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*(1), 22-24.

Carrow-Woofolk, E. (1985). *Test for Auditory Comprehension of Language- Revised Edition*. Austin, TX: Pro-Ed.

Coates, S., & Bromberg, P. M. (1973). Factorial structure of the Wechsler Preschool and Primary Scale of Intelligence between the ages of 4 and 6½. *Journal of Consulting and Clinical Psychology, 40*(3), 365-370.

CTB/McGraw Hill. (1992). *California Achievement Tests, Form E.* Monterey, CA: Author.

CTB/McGraw Hill. (1996). *Comprehensive Test of Basic Skills.* Monterey, CA: Author.

Das, J. P., Kirby, J. R. & Jarman, R. F., (1975). Simultaneous and successive syntheses: An alternative model for cognitive abilities. *Psychological Bulletin*, 82, 87-103.

Das, J. P., Kirby, J. R. & Jarman, R. F., (1979) Simultaneous and successive cognitive processes. New York: Academic Press.

Dunn, L., & Dunn, L. (1981). *Peabody Picture Vocabulary Test – Revised*. Circle Pines, MN: American Guidance Service.

Dunn, L., & Dunn, L. (1997). *Peabody Picture Vocabulary Test—Third Edition: Examiner's Manual*. Circle Pines, MI: American Guidance System.

Elliott, C.D. (1990). *Differential Ability Scales*. San Antonio, RX: The Psychological Corporation.

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.

Fenson, L., Dale, P., Reznick, S., Thal, D., Bates, E., Hartung, J., Pethick, S., & Reilly, J. (1993). *MacArthur Communicative Development Inventories: User's guide and technical manual*. San Diego, CA: Singular/Thomson Learning.

Ginsburg, H. P., & Baroody, A. J. (1990). *Test of Early Mathematics Ability, Second Edition: Examiner's manual*. Austin, TX: PRO-ED, Inc.

Glutting, J. J. (1986). Potthoff bias analyses of K-ABC MPC and Nonverbal Scale IQ's among Anglo, Black and Puerto Rican kindergarten children. *Professional School Psychology, 1*(4), 225-234.

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*(3), 237-248.

Hammill, D. D., Ammer, J. F., Cronin, M. E., Mandelbaum, L. H., & Quinby, S. S. (1987). *Quick-Score Achievement Test*. Austin, TX: Pro Ed.

Harms, T., Cryer, D., & Clifford, R. M. (1990). *Infant/Toddler Environment Rating Scale*. New York: Teachers College Press.

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*(2), 116-125.

Hresko, W.P., Reid, D.K., Hammill, D.D., Ginsburg, H.P., & Baroody, A.J. (1988). *Screening children for related early educational needs*. Austin, TX: Pro-Ed.

Huttenlocher, J., & Levine, S. C. (1990a). *Primary Test of Cognitive Skills: Examiner's manual*. Monterey, CA: CTB/McGraw Hill.

Huttenlocher, J., & Levine, S. C. (1990b). *Primary Test of Cognitive Skills: Norms book*. Monterey, CA: CTB/McGraw Hill.

Huttenlocher, J., & Levine, S. C. (1990c). *Primary Test of Cognitive Skills: Technical bulletin*. Monterey, CA: CTB/McGraw Hill.

Johnson, D. L., Howie, V. M., Owen, M., Baldwin, C. D., & Luttman, D. (1993). Assessment of three-year-olds with the Stanford-Binet Fourth Edition. *Psychological Reports, 73*(1), 51-57.

Kaufman, A. S., & Kaufman, N. L. (1983a). *Kaufman Assessment Battery for Children: Administration and scoring manual*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1983b). *Kaufman Assessment Battery for Children: Interpretive manual*. Circle Pines, MN: American Guidance Service.

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.

Laughlin, T. (1995). The school readiness composite of the Bracken Basic Concepts Scale as an intellectual screening instrument. *Journal of Psychoeducational Assessment 13*(3), 294-302.

LoBello, S. G. (1991). A short form of the Wechsler Preschool and Primary Scale of Intelligence-Revised. *Journal of School Psychology, 29*(3), 229-236.

Love, J. M., Kisker, E. E., Ross, C. M., Schochet, P. Z., Brookes-Gunn, J., Paulsell, D., Boller, K., Constantine, J., Vogel, C., Fuligni, A .S., & Brady-Smith, C. (2002). *Making a*

difference in lives of infants and toddlers and their families: The impacts of Early Head Start. Volume Final Technical Report.

Markwardt, F. C. (1998). *Peabody Individual Achievement Test—Revised: Manual* (Normative update). Circle Pines, MN: American Guidance Service.

Mather, N., & Woodcock, R. W. (2001a). *Woodcock-Johnson III Tests of Achievement: Examiner's manual*. Itasca, IL: The Riverside Publishing Company.

Mather, N., & Woodcock, R. W. (2001b). *Woodcock-Johnson III Tests of Cognitive Abilities: Examiner's manual*. Itasca, IL: The Riverside Publishing Company.

McCarthy, D. (1972). *McCarthy Scales of Children's Abilities*. San Antonio, TX: The Psychological Corporation.

McCormick, M. C., McCarton, C., Tonascia, J. & Brooks-Gunn, J. (1993). Early educational intervention for very low birth weight infants: Results from the Infant Health and Development Program. *Journal of Pediatrics, 123*(4), 527-533.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III: Technical manual*. Itasca, IL: The Riverside Publishing Company.

McGroder, S. M., Zaslow, M. J., Moore, K. A., & LeMenestrel, S. M. (2000). *National evaluation of welfare-to-work strategies. Impacts on young children and their families two years after enrollment: Findings from the Child Outcomes Study*. Washington, DC: Child Trends.

Newborg, J., Stock, J.R., Wnek, L. (1984). Battelle Developmental Inventory. Itasca, IL: Riverside Publishing.

NICHD Early Child Care Research Network (1999). Child outcomes when child care center classes meet recommended standards of quality. *American Journal of Public Health, 89*(7), 1072-1077.

NICHD Early Child Care Research Network (2000). The relation of child care to cognitive and language development. *Child Development, 71*(4), 960-980.

Peisner-Feinberg, E. S., & Burchinal, M. R. (1997). Relations between preschool children's child care experiences and concurrent development: The Cost, Quality, and Outcomes Study. *Merrill-Palmer Quarterly, 43*(3), 451-477.

Pierrehumbert, B., Ramstein, T., Karmaniola, A., & Halfon, O. (1996). Child care in the preschool years: Attachment, behaviour problems and cognitive development. *European Journal of Psychology of Education, 11*(2), 201-214.

Ramey, C. T., & Campbell, F. A. (1991). Poverty, early childhood education, and academic competence: The Abecedarian experiment. In A. C. Huston (Ed.), *Children reared in poverty: Child development and public policy* (pp. 190-221). New York: Cambridge University Press.

Ramey, C.T., Yeates, K.W., & Short, E. J (1984). The plasticity of intellectual development: Insights from preventative intervention. *Child Development, 55*, 1913-1925.

Saylor, C. F., Boyce, G. C., Peagler, S. M., Callahan, S. A. (2000). Brief report: Cautions against using the Stanford-Binet-IV to classify high-risk preschoolers. *Journal of Pediatric Psychology, 25*(3), 179-183.

Schneider, B. H., & Gervais, M. D. (1991). Identifying gifted kindergarten students with brief screening measures and the WPPSI-R. *Journal of Psychoeducational Assessment, 9*(3), 201-208.

Schweinhart, L.J., Barnes, H.V. & Weikart, D.P. (1993). *Significant benefits: The High/Scope Perry Preschool Study through age 27 (Monograph of the High/Scope Educational Research Foundation, 10)*. Ypsilanti, MI: High/Scope Press.

Slosson, R. L. (1983). *Intelligence Test (SIT) and Oral Reading Test (SORT): For Children and Adults*. Los Angeles: Western Psychological.

Tellegen, A., & Briggs, P. F. (1967). Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting Psychology, 31*(5), 499-506.

Terman, L.M. & Merrill, M.A. (1973). Stanford-Binet Intelligence Scale, Form LM. Itasca, IL: The Riverside Publishing Company.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986a). *Stanford-Binet Intelligence Scale: Fourth Edition. Guide for administering and scoring*. Itasca, IL: The Riverside Publishing Company.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986b). *Stanford-Binet Intelligence Scale: Fourth Edition. Technical manual*. Itasca, IL: The Riverside Publishing Company.

Wechsler, D. (1967). *Manual for the Wechsler Preschool & Primary Scale of Intelligence.* New York: The Psychological Corporation.

Wechsler, D. (1972). *Echelle d'intelligence de Wechsler pour la période préscolaire et primaire, W.P.P.S.I.* Paris, France: Les Editions du Centre de Psychologie Appliquée.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence – Revised (WPPSI-R): Manual*. San Antonio, TX: The Psychological Corporation, Harcourt Brace Jovanovich, Inc.

Weikart, D.P., Bond, J.T., and McNeil, J.T. (1978). *Ypsilanti Perry Preschool Project: Preschool Years and Longitudinal Results through Fourth Grade*. Ypsilanti, Mich.: High/Scope Press.

West, J. & Andreassen, C. (2002, May). *Measuring early development in the Early Childhood Longitudinal Study—Birth Cohort*. Paper prepared for Selecting Measures for Young Children in Large Scale Surveys, a workshop sponsored by the Research Network on Child and Family Well-Being and the Science and Ecology of Early Development, Washington, DC.

Williams, J. M., Voelker, S., & Ricciardi, P. W. (1995). Predictive validity of the K-ABC for exceptional preschoolers. *Psychology in the Schools, 32*(3), 178-185.

Woodcock, R.W. & Johnson, M.B. (1989). *Woodcock-Johnson Psycho-Educational Battery – Revised*. Itasca, IL: Riverside Publishing.

Zimmerman, I.L., Steiner, V.G., and Pond, R.E. (1992). *Preschool Language Scale-3 (PLS-3)*. San Antonio, TX: The Psychological Corporation.

Early Childhood Measures: Social-Emotional

Early Childhood Measures: Social-Emotional

Behavioral Assessment System for Children (BASC)

I. Background Information

Author/Source

Source: Reynolds, C.R., & Kamphaus, R.W. (1998). *BASC Behavioral Assessment System for Children: Manual*. Circle Pines, MN: American Guidance Service, Inc.

Publisher: American Guidance Service, Inc. (AGS)
4201 Woodland Rd.
Circle Pines, MN 55014-1797
Phone: 800-328-2560
Website: www.agsnet.com

Purpose of Measure

As described by instrument publisher:
The BASC was designed as a series of measures focusing on children's behavioral and emotional problems as well as positive behavioral and emotional characteristics. The BASC can be used to "…facilitate the differential diagnosis and educational classification of a variety of emotional and behavioral disorders of children and to aid in the design of treatment plans" (Reynolds & Kamphaus, 1998, p. 1). When used as a diagnostic tool, the BASC can help to distinguish children with severe emotional disturbances from children with less extreme problems, including conduct disorders and social maladjustment, as required by The Individuals with Disabilities Education Act. The manual indicates that the BASC "…can assess all aspects of the federal definition of severe emotional disturbance" (p. 6).

The authors also indicate usefulness for program evaluation and basic research on childhood psychopathology and behavior disorders.

The focus of this summary will be on the parent and teacher report forms for preschool and school-age children. The reader may refer to the published manual (Reynolds & Kamphaus, 1998) for information on measures available for use with older children and adolescents.

Population Measure Developed With

Two norming samples were used in the development of the BASC—a General sample and a Clinical sample. These samples will be described in the following sections. Collection of these samples took place from fall 1988 through spring 1991. In addition, because children under 4 years of age were not included in the original norming samples, a separate norming sample was used for children ages 2 years, 6 months to 3 years, 11 months . These data were collected from winter 1996 through spring 1998.

- *General norm samples*: The General samples were recruited from 116 sites across the United States. Sites were chosen so as to create samples that would be representative of the U.S. population of children ages 4 to 18 with respect to race/ethnicity, socioeconomic

status, and gender. Children with special needs enrolled in regular classrooms and preschool programs were also represented in the samples. Public and private schools and daycare centers were the primary testing sites. Additional settings, including PTA, church groups and health care centers were also used to recruit samples for the Parent Rating Scales.

- For the Teacher Rating Scales, 333 children ages 4 to 5 were included in the General norm sample for the Preschool version of the scale (TRS-P), and 1,259 children ages 6 to 11 were included in the sample for the Child version (TRS-C).
- For the Parent Rating Scales, 309 children ages 4 to 5 were included in the General norm sample for the Preschool version (PRS-P), and 2,084 children ages 6 to 11 were included for the Child version (PRS-C).
- Additional General and Clinical norming samples were recruited at a later time to obtain norms for 2 year, 6 month and 3-year-olds for the TRS-P and the PRS-P.
- Percentages of white, black, Hispanic, and "other minority" group children were represented in the General norm samples in approximately the same proportions as in the 1990 U.S. population estimates, with some exceptions:
  - Black and Hispanic children were somewhat overrepresented in the TRS-P sample;
  - Black children were also overrepresented in the TRS-C sample while Hispanic children were underrepresented; and
  - White children were overrepresented in the PRS-C sample.
  - Mothers who completed either preschool or child versions of the Parent Rating Scales tended to have higher than average levels of education compared with women ages 25 to 34 in the U.S. population.
- Weighting procedures were used to bring the samples into closer alignment with U.S. population estimates for race/ethnicity and mothers' education.

- *Clinical norm samples*: The Clinical sample was recruited from community mental health centers, public school classrooms, and programs for children with behavioral or emotional disturbances, residential programs for children with behavioral and emotional problems, university- and hospital-based inpatient and outpatient mental health services, and juvenile detention centers. Children in the General samples with diagnosed emotional or behavioral disorders were also included in the Clinical samples.
  - For the Teacher Rating Scales, 109 children ages 4 to 5 were included in the Clinical sample for the TRS-P, and 393 children ages 6 to 11 were included in the sample for the TRS-C.
  - For the Parent Rating Scales, 69 children ages 4 to 5 were included for the PRS-P, and 239 children ages 6 to 11 were included for the Child version PRS-C.
  - The most common diagnoses of children included in the Clinical norm samples were behavior disorder and attention deficit hyperactivity disorder (ADHD).
  - A large majority of children in the Clinical sample were white, ranging from 73 percent for the TRS-P to 90 percent for the PRS-P. Black representation ranged from a low of 3 percent for the PRS-P to a high of 20 percent for the TRS-P. Hispanic children constituted between 2 percent (TRS-C) and 6 percent (PRS-P) of the sample, and 1 percent to 2 percent of the sample was other minorities.

- *Young preschool norm samples*: As noted above, children under the age of 4 were not included in the original norm samples. Normative data for children ages 2 years, 6 months to 3 years, 11 months, were collected in winter 1996 through spring 1998. Only general norms were constructed, because diagnosis of clinical disorders occurs rarely among young preschoolers. Data were collected at forty-one sites across the U.S., primarily day care programs of varying types. The TRS-P was completed by day care staff who were very familiar with the children, and mothers completed the PRS-P. TRS-P forms were completed for 678 children, and PRS-P forms were completed for 637. However, some cases were dropped for each report, in order to bring the distribution of sex, race/ethnicity, and mother's education into closer alignment with U.S. population distributions. Ultimately, 664 children were included in the sample for TRS-P analyses, and 559 children were included for the PRS-P norming sample. Despite this, black children were substantially underrepresented in the PRS-P sample (8.6 percent compared with a 1994 U.S. population estimate of 16.1 percent of children ages 2-3), Hispanic children were underrepresented in both the PRS-P and TRS-P samples (9.3 percent and 10.4 percent, respectively, compared with a 15.2 percent population estimate), and white children were overrepresented in both samples (70.5 percent and 75.0 percent, compared with a 64.7 percent population estimate). Samples were subsequently weighted by race/ethnicity and (for the PRS-P sample) mothers' education, within gender.

Age Range Intended For

Ages 2 years, 6 months through 5 years (PRS-P and TRS-P), and ages 6 through 11 (PRS-C and TRS-C).

Key Constructs of Measure

There are 14 scales derived from the TRS-C and TRS-P, 12 of which are also included in the PRS-C and PRS-P. There are 10 clinical scales, tapping maladaptive behaviors, and 4 adaptive behavior scales. Several composite scores are derived from these scales.

- *Clinical Scales*:
  - *Aggression*: Verbally and physically aggressive actions toward others.
  - *Hyperactivity:* Tendencies toward overly high activity levels, acting without thinking, and rushing through work or activities.
  - *Conduct Problems*: Antisocial, noncompliant, and destructive behavior (TRS-C and PRS-C only).
  - *Anxiety:* Nervousness, fearfulness, and worries about real and imagined problems.
  - *Depression:* Includes sadness, moodiness, and low self-esteem.
  - *Somatization:* Complaints about relatively minor physical problems and discomforts.
  - *Attention Problems*: Distractibility and poor concentration.
  - *Learning problems:* Academic problems, particularly inability to adequately understand and complete schoolwork. (TRS-C only).
  - *Atypicality:* A collection of unusual, "odd" behaviors that may be associated with psychosis, such as experiencing visual or verbal hallucinations and self-injurious behavior.
  - *Withdrawal:* Avoidance of social contact.

- *Adaptive Behavior Scales*:
  - *Adaptability:* The ability to adjust to changes in the environment.
  - *Leadership:* Includes the ability to work well with others, social activity, and creativity (TRS-C and PRS-C only).
  - S*ocial Skills:* Behaviors that facilitate positive interactions with peers and adults.
  - *Study Skills:* Good study habits (TRS-C only).
- *Composites*:
  - *Externalizing Problems:* Includes the Aggression and Hyperactivity scales, as well as Conduct Problems scale for child and adolescent levels.
  - *Internalizing Problems:* Includes the Anxiety, Depression, and Somatization scales.
  - *School Problems:* For the TRS child and adolescent only, summarizes the Attention Problems and Learning Problems scales.
  - *Adaptive Skills:* Consists of Adaptability, Social Skills, and Leadership scales, as well as the Study Skills scale for the teacher report. Composition varies by age level, as not all scales are included for all three levels.
  - *Behavioral Symptoms Index:* Includes the Aggression, Hyperactivity, Anxiety, Depression, Attention Problems, and Atypicality scales.

Norming of Measure (Criterion or Norm Referenced)

Norm referenced. Norms have been established for individual scales as well as for the composite scores. There are separate General and Clinical norms (ages 4 and above only), and Female and Male norms are also available.

Comments

- Reynolds and Kamphaus (1998) present extensive information on how preliminary sets of items and scales were developed, tested, modified and retested in order to produce the final item and scale structures for each of the BASC measures. The procedures used in measure construction appear to have been both extensive and rigorous, and equal rigor went into the subsequent norms development.
- Although this measure does include positive behavior scales, it is primarily a diagnostic tool and is heavily weighted toward detecting behavioral and emotional problems. The importance of adaptive behaviors is discussed by Reynolds and Kamphaus (1998) as facilitating understanding of children's strengths that should be considered when developing individualized educational and treatment plans.
- The BASC is a relatively new set of measures. The scales are highly clinical, and most of the research that has been conducted with BASC measures has focused on differential diagnosis of behavioral disorders. Little research has been conducted as of yet addressing usefulness of the BASC for other purposes, such as examinations of the extent to which children's adjustment as assessed with BASC measures is modifiable through changes in a classroom environment, the meaningfulness of describing classrooms and other groups of children with average scores on scales and composites, and the extent to which individual variations in scores within a normal range are predictive of subsequent positive or negative outcomes. However, findings of expectable associations between teacher ratings and children's standardized math and reading test performance (Merydith, 2001, described below) is promising in this regard, and the BASC has received positive

evaluations of its usefulness for assessment of children's behavioral and emotional problems, and in particular for school-based assessments (e.g. Flanagan, 1995).

II. Administration of Measure

Who is the Respondent to the Measure?

- Parent. Parents or guardians complete the PRS-P and PRS-C.
- Teacher. Teachers or other adults complete the TRS-P and TRS-C. Reynolds and Kamphaus (1998) indicate that respondents should have had a month of daily contact with the child or children they are evaluating, or six to eight weeks of contact several days a week. The authors further suggest that it is preferable for adults completing the TRS-C (for school-age children) to have supervised the child or children being evaluated in structured classroom settings.

If Child is Respondent, What is Child Asked to Do?

N/A.

Who Administers Measure/Training Required?

Test Administration:

- The PRS-P, PRS-C, TRS-P, and TRS-C are questionnaires that are typically given to parents and teachers to complete independently, although the PRS-P and -C have also been administered in an interview format.

Data Interpretation:

- Little training is required for administration or scoring, although Reynolds and Kamphaus (1998) caution that "…interpreting and applying the results require a level of sophistication in psychology and psychological and educational testing typically obtained through graduate education…" (Reynolds & Kamphaus, 1998, p. iv).

Setting (e.g., one-on-one, group, etc.)

Parents and teachers usually complete rating scales on their own, preferably in a single sitting in a location where distractions are minimized. There is an online administration option available from the publisher for the parent and teacher reports as well as for other BASC measures.

Time Needed and Cost

Time:

- Both the PRS and the TRS take between 10 and 20 minutes to complete.

Cost:

- BASC ASSIST Scannable Windows Kit: $482.95
- Manual: $76.95
- PRS and TRS forms: $38.95 per pkg. of 25 scannable forms, or $29.95 per pkg. of hand-scored forms.

- There are numerous other options for purchasing BASC materials, including several options for scoring software and services, including online administration and scoring services.

Comments

- This measure has been designed to be easily understood by parents and teachers and to take a fairly short time to complete.

III. Functioning of Measure

Reliability Information from Manual

In the analyses described below, reliability estimates for the PRS-P and TRS-P refer to analyses conducted with 4- to 5-year old children (the preschool sample in the original norm samples) unless otherwise noted.

Internal consistency of Teacher Report Scales

- For the teacher report, coefficient alphas for the General norm sample ranged from .78 to .90 for the TRS-P scales (ages 4 and 5), with a median alpha of .82. For the TRS-C in the General sample, internal consistencies were assessed separately for younger (ages 6 to 7) and older (ages 8 to 11) children. At the younger ages, coefficient alphas ranged from .62 to .94, with a median of .84. Alphas for the older age group ranged from .77 to .95, with a median alpha of .88. Alphas for the composites for the TRS-P and the TRS-C (both younger and older age groups) ranged from .89 to .97 (see Reynolds, & Kamphaus, 1998, p. 102).
- Internal consistencies of the TRS-P in the norm sample for younger preschoolers (2 years, 6 months to 3 yrs., 11 months) ranged from .71 to .92, with a median of .80 for the individual scales, and from .89 to .95 for the composites (see Reynolds, & Kamphaus, 1998, p. 305).
- In the Clinical norm samples, alphas for the TRS-P scales ranged from .66 to .91, with a median of .84. Alphas for the TRS-C ranged from .74 to .94, with a median of .82. Internal consistency of the composites ranged from .82 to .94 for the TRS-P and from .89 to .95 for the TRS-C (see Reynolds, & Kamphaus, 1998, p. 103).

Internal consistencies of Parent Report Scales

- For the parent report, General norm samples alphas ranged from .69 to .86 for the PRS-P scales, with a median of .74. As with the TRS-C, internal consistencies of PRS-C scales and composites in the General sample were assessed separately for younger and older children. At the younger ages, coefficient alphas ranged from .51 to .89, with a median of .80. Alphas for the older age group ranged from .58 to .89, with a median alpha of .78. The .51 and .58 alphas at the two ages were for the same scale, Atypicality, and at both ages these alphas were much lower than the second lowest alphas, .67 for Somatization at the younger age and .71 for Conduct Problems at the older age. Alphas for the composites across the two age groups ranged from .86 to .93 (see Reynolds, & Kamphaus, 1998, p. 130).

- Internal consistencies of the PRS-P in the norm sample for younger preschoolers were established separately for children under age 3 and children between 3 and 4 years of age. Coefficient alphas were similar at both ages, ranging from .59 to .84 for the individual scales (median reliabilities of .69 and .75 for the younger and older age groups, respectively), and from .82 to .91 for the composites (see Reynolds, & Kamphaus, 1998, p. 305).
- In the Clinical norm sample alphas for the PRS-P scales ranged from .72 to .91, with a median of .83. Alphas for the PRS-C ranged from .69 to .89, with a median of .80. Internal consistency of the composites was similar for the two age levels, ranging from .84 to .94 (see Reynolds, & Kamphaus, 1998, p. 131).

Test-retest reliability of Teacher Ratings

- A subsample of children from both the General and the Clinical norm samples were evaluated twice by their teachers, with an interval ranging from 2 to 8 weeks between ratings. For the TRS-P, correlations ranged from .78 to .93 for the scales, and from .83 to .95 for the composites, with a median correlation of .89. For the TRS-C, correlations ranged from .70 to .94 for the scales, and from .85 to .95 for the composites. The median correlation was .91 (see Reynolds, & Kamphaus, 1998, p. 105).
- The longer-term stability of TRS-C ratings was also examined for a sample of behaviorally disordered and emotionally disturbed children (all white, 75 percent male) from one school district. These children were rated by their teacher a second time, 7 months after the initial TRS-C administration. Correlations ranged from .37 to .78 for scales and from .58 to .76 for composites, with a median correlation of .69 (see Reynolds, & Kamphaus, 1998, p. 108).

Test-retest reliability of Parent Ratings

- Test-retest reliabilities for the PRS-P and PRS-C were established in small samples of children drawn from both the General and Clinical norm samples. Each child was rated twice by the same parent, with an interval of 2 to 8 weeks between ratings. For the PRS-P, correlations ranged from .61 to .91 for individual scales, and from .79 to .88 for composites, with a median correlation of .85. Test-retest correlations of PRS-C scales ranged from .71 to .91 for scales, and from .83 to .92 for the composites, with a median correlation of .88 (see Reynolds, & Kamphaus, 1998, p. 132).

Interrater reliability of Teacher Ratings

Two forms of interrater reliability were presented, both involving agreement between ratings by teachers on the TRS-P or the TRS-C.

- The first form, available only for the TRS-P, utilized interrater correlations of four pairs of teacher raters. Each pair of teachers rated between 8 and 20 children, and scale and composite scores based on these two ratings were correlated. The overall interrater reliability estimates were then reported as weighted averages of four resulting correlations (one for each pair of teacher raters). As described by Reynolds and Kamphaus (1998), "…these interrater correlations represent the degree to which teachers rank children in the same way on each dimension of behavior" (p. 104). Interrater correlations for Somatization and Adaptability were .27 and .38, respectively, while

correlations for other scales ranged from .50 to .76. Correlations for composites ranged from .63 to .69 (see Reynolds, & Kamphaus, 1998, p. 106).

- The second form of interrater reliability was calculated for both preschool and child age levels. Data from many pairs of teachers, each pair of whom may have rated only one child in common, are combined so that one member of each pair is randomly assigned to be Rater 1, and the other to be Rater 2. Correlations between Rater 1 and Rater 2 constitute the measure of interrater reliability. According to Reynolds and Kamphaus (1998), "…this type of data reflects the degree to which ratings from different teachers are interchangeable; that is, it reflects agreement both in the rank ordering of children and in the level of scores assigned" (p. 104). The interrater correlation for the TRS-P Somatization scale was .27. The remaining correlations for the TRS-P scales ranged from .49 to .69, and correlations for the composites ranged from .43 to .72. For TRS-C scales the range of correlations was .53 to .94, and interrater correlations for composites ranged from .67 to .89 (see Reynolds, & Kamphaus, 1998, p. 106).

Interrater reliability of Parent Ratings

- Interrater reliability for parent reports were examined in small samples of preschool and elementary school-age children who were rated by both mothers and fathers. Because each set of parents rated only one child (their own), inter-parent reliability can be interpreted in the same manner as the second form of inter-teacher reliability described above. Inter-parent correlations for the PRS-P ranged from .34 to .59 for individual scales, and from .40 to .57 for composites, with a median scale reliability of .46. For PRS-C scales and composites, correlations ranged from .30 to .73 for scales, and from .47 to .78 for the composites. The median correlation for the PRS-C was .57 (see Reynolds, & Kamphaus, 1998, p. 134).
- In a small sample of younger preschoolers, inter-parent correlations ranged from .36 to .66 for individual scales, and from .47 to .65 for composites, with a median scale reliability of .59. This compares favorably with the .46 correlation found for older preschoolers (see Reynolds, & Kamphaus, 1998, p. 306).

Validity Information from the Manual

Construct validity of Teacher Report Scales

To examine the construct validity of the BASC composites, Reynolds and Kamphaus (1998) reported two different types of factor analysis of data from the General norm samples. The first of these was covariance structure analysis (CSA), in which the expected factor model is assessed to determine how well it fits the actual questionnaire response patterns (a form of confirmatory factor analysis). The second type was principal axis factoring, in which no *a priori* model is tested but rather a model is created that optimally fits the data. The Behavioral Symptoms Index was not investigated in these analyses, but the Withdrawal scale, which is not included in any composite, the Attention Problems scale, which is not part of a composite at the preschool level, and the Atypicality scale, which is included only in the Behavioral Symptoms Index, were included.

- As discussed by Reynolds and Kamphaus (1998, pp. 111-117), results of analyses for the TRS-P and the TRS-C were generally supportive of the 3 composites of the TRS-P and the 4 composites of the TRS-C, although the results also indicated that Externalizing

Problems, Internalizing Problems, Adaptive Skills, and School Problems (TRS-C only) as assessed with the BASC are not independent of each other.

- As reported by Reynolds and Kamphaus (pp. 114-116) the Attention Problems scale had negative cross-loadings on an Adaptive Skills factor for the TRS-P (-.56 in CSA; -.64 in principal axis analyses) and for the TRS-C (-.56 in principal axis analyses only). Learning Problems had a similar negative cross-loading (-.46) on an Adaptive Skills factor in principal axis analyses of the TRS-C. The inclusion of Atypicality on the Behavioral Symptoms Index but on neither the Internalizing nor the Externalizing composites received some support from its approximately equal associations with Internalizing and Externalizing factors at both age levels (.43 and .42 for CSA of the TRS-P; .41 and .44 for CSA of the TRS-C; .50 and .48 for principal axis analyses of the TRS-P; .46 and .47 for principal axis analyses of the TRS-C).
- Subsequent analyses with TRS-P data from younger preschoolers indicated few differences in the functioning of the composites between the younger and older preschool groups.

Construct validity of Parent Report Scales

- As reported by Reynolds and Kamphaus (1998, 139-143), results of analyses for the PRS-P and the PRS-C supported the presence of 3 factors at both the preschool and elementary school levels, reflecting Externalizing Problems, Internalizing Problems, and Adaptive Skills. Also consistent with teacher-report findings were indications of the interrelations among behaviors tapped by the scales and composites. Depression, which is part of the Internalizing Problems composite, displayed cross-loadings on Externalizing Problems factors for both the PRS-P and the PRS-C (.49 for CSA of the PRS-P; .49 for CSA of the PRS-C; .49 for principal axis analyses of the PRS-P; and .45 for principal axis analyses of the PRS-C). In addition, Adaptability loaded primarily on the Adaptive Skills factor in analyses of the PRS-C, but also had a negative cross-loading of -.42 on the Externalizing Behavior Problems factor in principal axis analyses.
- Additional factor analyses of PRS-P data for younger preschoolers produced results that were almost identical to results with the older preschoolers.

Convergent validity of Teacher Report Scales

Reynolds and Kamphaus (1998) conducted several studies investigating associations between TRS-P and TRS-C ratings and ratings on other measures tapping behavior problems and adaptive behavior, including the Child Behavior Checklist - Teacher's Report Form (CBCL-TRF; Achenbach, 1991), Conners' Teacher Rating Scales (CTRS-39; Conners, 1989), Burks' Behavior Rating Scales (BBRS; Burks, 1977), and the Teacher Rating Scale of the Behavior Rating Profile (BRP; Brown & Hammill, 1983). Of these four, the study including the CTRS-39 was conducted with a preschool sample, while the remaining three involved ratings of elementary school-age children. Results from all of these studies found associations between scales and composites tapping similar constructs.

- Associations between CBCL-TRF and TRS-C scales and composites tapping similar constructs included correlations of .88 for Externalizing and .73 for Internalizing, and the TRS-C Behavioral Symptoms Index correlated .92 with the CBCL-TRF Total Problems composite. Although the TRS-C Adaptive Skills composite differs considerably in content from the CBCL-TRF Total Adaptive Functioning composite, these two indicators

of positive functioning correlated .75. School Problems, which does not have a directly comparable composite on the CBCL-TRF, correlated .74 with the Total Problems composite (see Reynolds, & Kamphaus, 1998, p. 118-119).

- Associations between BBRS and TRS-C scales and composites tapping similar constructs included correlations ranging from .79 to .89 between the TRS-C Externalizing Problems composite and BBRS scales tapping poor control of impulses and anger, aggressiveness, and noncompliance. The TRS-C Internalizing Problems composite correlated .73 and .74 with BBRS scales tapping self-blaming and anxiety. The TRS-C School Problems Composite demonstrated correlations ranging from .66 to .94 with the BBRS scales reflecting intellectual, academic, and attentional problems; and the Behavior Problems Index had correlations ranging from .36 to .88 with all BBRS scales, with a median correlation of .69. There are no positive behavior scales on the BBRS, but the TRS-C Adaptive Skills composite was correlated -.33 to -.87 with all of the BBRS scales, with a median correlation of -.67 (see Reynolds, & Kamphaus, 1998, p. 123).
- The BRP yields a single profile score, with lower scores reflecting more negative behaviors. All correlations were in the expected direction; adaptive behaviors from the TRS-C were positively correlated with BRP scores and problem behaviors were negatively correlated with BRP scores. Correlations between the TRS-C and the BRP ranged in absolute value from .24 for Withdrawal to .60 for Learning Problems and Behavioral Symptoms Index ratings (see Reynolds, & Kamphaus, 1998, p. 124).
- Associations between TRS-P Externalizing Problems composite and Behavioral Symptoms Index scores with CTRS-39 Hyperactivity, Conduct Problems, and Hyperactivity Index scores ranged from .60 to .69. Other correlations of similar magnitude were found between TRS-P Aggression scale scores and CTRS-39 Hyperactivity and Conduct Problems scores (.61 and .63, respectively), and between TRS-P Depression ratings and CTRS-39 Emotional Overindulgent ratings (.69). There are no positive scales on the CTRS-39; the TRS-P Adaptive Skills composite correlated -.14 to -.49 with the CTRS-39 scales (see Reynolds, & Kamphaus, 1998, p. 121).

Convergent validity of Parent Report Scales

Reynolds and Kamphaus (1998) also reported studies investigating associations between parent ratings and ratings on other measures, including the parent-report Child Behavior Checklist (CBCL; Achenbach, 1991), Conners' Parent Rating Scales (CPRS-93; Conners, 1989), the Personality Inventory for Children-Revised (PIC-R; Lachar, 1982), and the Parent Report Form of the Behavior Rating Profile (BRP; Brown & Hammill, 1983). Preschool samples were used in studies with the CBCL and the PIC-R, and studies with elementary school-age children were conducted with the CBCL, the CPRS-93, and the BRP. As with the studies involving teacher reports, all of these studies with parent ratings indicated expectable associations between scales and composites tapping the same or similar constructs.

- The PRS-P Externalizing Problems composite was correlated .79 with the CBCL Externalizing composite, and was also correlated .58 with the CBCL Internalizing composite. The PRS-P Internalizing Problems composite was correlated .65 with both the Internalizing and Externalizing composites from the CBCL. The PRS-P Behavioral Symptoms Index correlated .86 with the CBCL Total Problems composite (see Reynolds, & Kamphaus, 1998, p. 144).

- The PRS-C Externalizing Problems composite was correlated .84 with the CBCL Externalizing composite, while the correlation with the CBCL Internalizing composite was only .33. The PRS-C Internalizing Problems composite was correlated .67 with the Internalizing composite from the CBCL, but only .23 with the CBCL Externalizing composite. The PRS-C Behavioral Symptoms Index correlated .81 with the CBCL Total Problems composite. The PRS-C Adaptive Skills composite was correlated .68 with CBCL Total Competence scores (see Reynolds, & Kamphaus, 1998, p. 145).
- Correlations between PRS-P scales and similarly-named PIC-R scales ranged in absolute value from .12 (PRS-P Somatization and Hyperactivity with PIC-R Somatic Concern and Hyperactivity, respectively) to .57 (PRS-P and PIC-R Withdrawal; see Reynolds & Kamphaus, 1998, p. 148). Reynolds and Kamphaus (p. 147) suggest that correlations may have been relatively low in some cases due in part to the inappropriateness of some items from the PIC-R for preschool children (e.g., questions pertaining to smoking, delinquent behavior, school and extracurricular activities).
- Associations between PRS-C and CPRS-93 were somewhat higher for scales tapping externalizing symptoms than for those tapping internalizing symptoms across the two measures. The PRS-C Externalizing Problems composite was correlated .78 with the CPRS-93 Conduct Disorder scale, .71 with the Antisocial scale, and also .67 with the Learning Problems scale. In contrast, the highest correlation of the PRS-C Internalizing composite with a CPRS-93 scale was .51 with Anxious-Shy. There are no positive behavioral scales on the CPRS-93. The PRS-C Adaptive Skills composite correlations with CPRS-93 scores ranged from -.48 with CPRS-93 Anxious-Shy to .07 with CPRS-93 Obsessive-Compulsive (see Reynolds, & Kamphaus, 1998, p. 149).

Reliability/Validity Information from Other Studies

- Merydith (2001) provided both reliability and validity information for the TRS-P, TRS-C, PRS-P, and PRS-C measures from a study of children of differing racial/ethnic groups enrolled in 12 kindergarten and first grade classrooms.
    - Temporal stabilities of TRS-P and TRS-C scales and composites across a 6-month time span were consistent with those reported by Reynolds and Kamphaus (1998) in their sample of behaviorally disordered and emotionally disturbed children. Merydith found correlations ranging from .12 to .76, with a mean correlation of .47. Correlations ranged from .48 to .68 for composites.
    - Merydith correlated TRS-P and TRS-C Internalizing, Externalizing, School Problems, Behavioral Symptoms Index, and Adaptive Skills composites and the Hyperactivity scale with parallel scales from the Social Skills Rating System (SSRS; Gresham & Elliott, 1990). Correlations ranged from .60 to .88.
    - TRS (-P or –C) Externalizing scores were significantly more highly correlated with SSRS Externalizing ratings than with SSRS Internalizing ratings, and TRS Internalizing scores were significantly more highly correlated with SSRS Internalizing ratings than with SSRS Externalizing ratings. According to Merydith, these findings provide support for the discriminant validity of the Externalizing and Internalizing composites.
    - Correlations across parallel scales from the PRS (-P or -C) and parent reports on the SSRS (Gresham & Elliott, 1990) were significant but somewhat lower than correlations across teacher reports, ranging from .49 to .72.

- As with the teacher report findings, PRS (–P or –C) Externalizing was significantly more highly correlated with the SSRS Externalizing than with SSRS Internalizing, and PRS Internalizing was significantly more highly correlated with SSRS Internalizing than with SSRS Externalizing.
- TRS (-P or –C) Learning Problems scores were significantly negatively correlated with children's math and reading scores from standardized achievement tests (-.44 and -.41, for math and reading, respectively), and the TRS Attention Problems scale correlated -.33 with children's standardized math scores.

- Flanagan, Alfonso, Primavera, Povall, and Higgins (1996) also examined associations between TRS-P and PRS-P ratings and SSRS teacher ratings in a small sample of predominantly black kindergartners attending a parochial school in a high-poverty community. These researchers reported correlations of TRS-P and PRS-P scales and composites with SSRS Social Skills and Problem Behaviors scales only.
  - Correlations between the TRS-P scales and the SSRS scales were considerably lower than those reported by Merydith (2001). The Adaptive Skills composite correlated .37 with the SSRS Social Skills scale. The Social Skills scale had a nonsignificant correlation of .22 with the SSRS Social Skills scale. The Behavioral Symptoms Index had a correlation of .60 with the SSRS Problem Behaviors scale.
  - Flanagan and colleagues found a significant correlation of .32 between the PRS-P Behavioral Symptoms Index and the SSRS Problem Behaviors scale, and higher significant correlations of .62 and .58 between the SSRS Social Skills scale and the PRS-P Adaptive Skills Composite and Social Skills scale, respectively.

Comments

- Internal consistency estimates were high at all ages for composites derived from the PRS-P, PRS-C, TRS-P, and TRS-C (i.e., Externalizing Problems, Internalizing Problems, School Problems, Adaptive Skills, and the Behavioral Symptoms Index). Internal consistencies reported for some of the individual scales from the BASC measures (both teacher- and parent-report measures) were moderate (between .60 and .69) or low (below .60). On the PRS-P, the median coefficient alpha for individual scales was somewhat lower for the youngest preschool age group than for older age groups, possibly indicating that internal consistency of parent reports may be somewhat lower for younger children than for older children.
- Test-retest correlations over a short time interval (2 to 8 weeks) were high for all scales and composites of the TRS-P, TRS-C, PRS-P, and PRS-C, providing support for the reliability of these measures. Further, test-retest correlations for the TRS-C across a seven month interval, although predictably lower than across the shorter intervals, remained high for composites, and moderate to high for individual scales, in a sample of children with identified emotional and behavioral problems. These findings provide further support for the reliability of the TRS-C (as well as evidence of some stability in children's behavior across time).
- With respect to interrater reliability, results reported by Reynolds and Kamphaus (1998) suggest a moderate degree of agreement across teacher ratings (with correlations for the Somatization scale falling in low range on the TRS-P, and moderate to high correlations for other TRS-P and TRS-C scales and composites), and stronger agreement for ratings of elementary school-age children than for ratings of preschool children. Both methods of

estimating interrater reliability appear to indicate that Somatization may be particularly difficult to rate reliably with preschoolers.

- For both the PRS-P and the PRS-C, correlations between mother- and father-ratings were moderate to high for both scales and composites. Overall, these correlations suggest a reasonable amount of consistency in the ways that mothers and fathers perceive and rate their children's behavior, but substantial differences as well.
- Overall, information provided by Reynolds and Kamphaus (1998) as well as information provided in independent reports by Merydith (2001) and Flanagan and colleagues (1996) supports the validity of BASC scales and composites. For preschool children, parent reports of externalizing problems and internalizing problems on the CBCL and the PRS-P were all highly interrelated, suggesting that at this age, children who are perceived by their parents as being relatively high in one type of problem are likely to be perceived as being relatively high in the other type of problem as well. As discussed in our profile of the CBCL/1½-5 and C-TRF, CBCL Internalizing and Externalizing scales tend to be highly correlated in general (i.e., nonclinical) populations of preschool children (see Achenbach & Rescorla, 2000), and thus these high intercorrelations for externalizing scales with internalizing scales across measures may reflect as much or more on the CBCL as on the BASC.
- Reasons for the discrepancies between reports of associations between BASC and SSRS scales in studies by Merydith (2001) and Flanagan and colleagues (1996) are unclear. The children in the Flanagan study were drawn from only two classrooms, and all ratings were conducted by only two teachers, while 12 classrooms were involved in the Merydith study. It may be that individual teacher characteristics may have had a strong influence on the results from the Flanagan study. The differences in ethnic and socioeconomic make-up of the two samples also may have been a source of variability across the two studies.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

None found.

V. Adaptations of Measure

Spanish Version of the BASC Parent Rating Scales

Description of Adaptation

The Spanish-language versions of the PRS-P, the PRS-C, and the PRS-A were developed through a process of having several bilingual English-Spanish experts review proposed items and suggest modifications. No back-translation process was reported.

Psychometrics of Adaptation

No psychometrics were reported by Reynolds and Kamphaus (1998).

Study Using Adaptation

None found.

Early Childhood Measures: Social-Emotional

Early Childhood Measures: Social-Emotional

Child Behavior Checklist/1½ -5 (CBCL/1½-5) and Caregiver-Teacher Report Form (C-TRF)

I. Background Information

Author/Source

Authors: Developed by T.M. Achenbach. Newest revised versions of the measures and manuals are co-authored by L.A. Rescorla (Achenbach & Rescorla, 2000; Achenbach & Rescorla, 2001).

Publisher: University of Vermont, Research Center for Children, Youth, & Families
Manuals and other materials available from:
ASEBA (Achenbach System of Empirically Based Assessment)
1 South Prospect St.
Burlington, VT 05401-3456
Telephone: 802-656-8313
Website: www.aseba.org

Purpose of Measure

The Child Behavior Checklist/ 1½ - 5 (CBCL/1½-5), and the Caregiver-Teacher Report Form (C-TRF) for ages 1 years, 6 months through 5 years, are components of the Achenbach System of Empirically Based Assessment (ASEBA), which also includes measures for assessment of older children, adolescents, and young adults.

As described by instrument publisher or author:
ASEBA measures are clinical instruments primarily designed to assess behavioral and emotional problems in children and adolescents. Achenbach and Rescorla (2000) indicate that there are both practical and research applications of the ASEBA measures. Practical applications include using ASEBA measures in educational settings to identify problems that individual children may have, to suggest the need for additional evaluation, to guide the development of individualized intervention plans, and to track changes in functioning. ASEBA measures can be used as both predictors and outcomes in basic and applied developmental and child clinical research. ASEBA measures can be useful in research investigating, for example, developmental changes in behavioral and emotional disorders, impacts of early-appearing behavioral and emotional problems on children's social and emotional development, effects of different conditions in children's physical and social environments on mental health outcomes, and effects of interventions on children's behavioral and emotional functioning.

Population Measure Developed With

ASEBA measures have been recently revised and renormed, in part with the purpose of allowing a single measure to be used across the preschool years. Formerly, there was a parent or caregiver report for ages 2-3 (the CBCL/2-3) and parent and teacher reports for ages 4-18 (the CBCL/4-18 and the TRF).

- *CBCL/1½-5 Samples*: Two overlapping samples were used for different purposes in the development of the CBCL/1½-5—a normative sample, and a higher risk sample used for factor analyses and development of scales.
    - The normative sample was derived from a national probability sample (the National Survey) collected in 1999 by the Institute for Survey Research. Preschoolers who had received mental health or special education services were excluded from the normative sample. A total of 700 nonreferred children (362 boys, 338 girls) were included. The majority (56 percent) of the children were white, 21 percent were black, 13 percent were Hispanic, and 10 percent were identified as mixed or other. Seventy-six percent of the CBCL/1½-5 respondents were mothers, 22 percent were fathers, and 2 percent were others. Socioeconomically, 33 percent of the sample was classified as upper SES, 49 percent was middle SES, and 17 percent was lower SES.
    - The second sample was designed to include children with relatively high levels of parent-reported behavior problems. It included children from the National Survey who were excluded from the normative sample due to receipt of mental health or special education services, children included in the normative sample whose CBCL/1½-5 Total Problems scores were at or above the median for the sample, and additional children from 5 other general population samples and 19 clinic settings whose Total Problems scores were at or above the normative sample median. A total of 1,728 children (922 boys, 806 girls) from diverse socioeconomic backgrounds were included, 59 percent white, 17 percent black, 9 percent Hispanic, and 15 percent mixed or other. Scales for the new version of the preschool parent-report, the CBCL/1½-5, were constructed with data from these 1,728 children. This sample of children exhibiting relatively high levels of problem behaviors was used in factor analyses for establishing the syndromes. Mothers completed 88 percent of the forms for this sample, fathers completed 10 percent, and 2 percent were completed by others.
- *C-TRF Samples*: As with the CBCL/1½-5, two separate samples were used for development of the C-TRF—a normative sample and a second sample of children with relatively high levels of teacher-rated behavioral and emotional problems.
    - The normative sample for the C-TRF included a total of 1,192 children (588 boys, 604 girls). Of these, 203 (95 boys, 108 girls) were children who were also part of the normative sample for the CBCL/1½-5 (and whose parents gave consent to contact a day care provider or teacher). The sample also included 989 children who had been part of a previous (1997) C-TRF norming sample, 753 of whom were participants in the NICHD Study of Early Child Care. The remaining children were drawn from 14 daycare centers and preschools located in 12 different states. In this sample, 48 percent of children were white, 36 percent were black, 8 percent were Hispanic, and 9 percent were mixed or other. This sample was more skewed to higher SES than was the C-PRF normative sample, with 47 percent classified as upper SES, 43 percent middle SES, and 10 percent lower SES.
    - The second sample included children from the National Survey sample whose C-TRF Total Problems scores were at or above the median for the sample. Also

included were additional children whose Total Problems scores were at or above the normative sample median, obtained from 7 other general population samples and 11 clinic settings. A total of 1,113 children were included (675 boys, 438 girls), 68 percent white, 20 percent black, 4 percent Hispanic, and 8 percent mixed or other, from diverse socioeconomic backgrounds.

Age Range Intended For

Children ages 1 year, 6 months through 5 years, 11 months.

Key Constructs of Measure

There are six factor-analytically derived "syndromes" that are consistent across parent and teacher preschool assessments (the CBCL/1½-5 and the C-TRF), and an additional syndrome assessed only with the CBCL/1½-5. There are also three summary scales from each measure, as well as an alternative scoring system oriented around diagnostic categories found in the Diagnostic and Statistical Manual of the American Psychiatric Association (DSM-IV; American Psychiatric Association, 1994).

- Syndromes:
  - *Emotionally Reactive*: General negative emotionality, moodiness, and problems adapting to change.
  - *Anxious/Depressed*: Clinginess, sensitivity, sadness, fearfulness and self-consciousness.
  - *Somatic Complaints*: Headaches, nausea, other aches and pains, and excessive neatness.
  - *Withdrawn*: Immaturity, low social responsiveness, apathy.
  - *Attention Problems*: Poor concentration, inability to stay on-task, excessive movement.
  - *Aggressive Behavior*: Anger, noncompliance, destructiveness, physical and verbal aggression towards others.
  - *Sleep Problems* (CBCL/1½-5 only): Trouble sleeping, nightmares, resistance to sleep, frequent waking.
- Summary scales:
  - *Internalizing*: Summarizes Emotionally Reactive, Anxious/Depressed, Somatic Complaints, and Withdrawn
  - *Externalizing*: Summarizes Attention Problems and Aggressive Behavior
  - *Total Problems*: A summary score of all problems items
    - For the CBCL/1½-5, this includes Sleep Problems items and other problem items that are not part of any scale, including one parent-identified problem (a problem not already listed among the CBCL/1½-5 items that the parent records and then rates in the same manner as the listed items).
    - For the C-TRF, this includes standard problem items not included on any scale, and one teacher-identified problem not included among the standard items.
- DSM-Oriented scales:

- *Affective Problems*: Negative affect, eating and sleeping disturbances, underactivity.
- *Anxiety Problems*: Clinginess, fearfulness.
- *Pervasive Developmental Problems*: Inability to adapt to change, lack of social responsiveness, rocking, speech problems, strange behavior.
- *Attention Deficit/Hyperactivity Problems*: Concentration problems, excessive movement, inability to tolerate delay, disruptive activity.
- *Oppositional Defiant Problems*: Anger and noncompliance.

Norming of Measure (Criterion or Norm Referenced)

Norm referenced (although raw scores which are neither norm nor criterion referenced are used most frequently in research applications).

Comments

- The versions of the CBCL and TRF described here are revisions of earlier measures. The CBCL/1½-5 is a revision of the CBCL/2-3 (Achenbach, 1992). The C-TRF is a revision of a C-TRF for 2 to 5 year old preschoolers (Achenbach, 1997). In both cases, revisions to the measures involved minor changes in wording and, in addition, two CBCL/2-3 items that were not included on any scale were replaced entirely. The creation of the form for the expanded preschool range should be a substantial benefit to educators and psychological professionals wishing to track consistency and change in children's behavioral and emotional adjustment across the preschool years.
- The CBCL and TRF are among the most well-known and widely used instruments in developmental and child clinical psychology research. A great deal of information is available relevant to usefulness with varying populations.
- The focus of all ASEBA measures is almost entirely on emotional and behavioral problems. There are no competence or strengths measures included in the preschool measures. A second measure would thus be required to tap positive behavioral and emotional characteristics.
- Although the length of time that is required to complete the CBCL/1½-5 does not appear to be any longer than for most other measures reviewed, there may be excessive redundancy in some areas—particularly the Aggressive Behavior scale, which is reported to have internal consistencies in excess of .90 and which includes 19 items in the CBCL/1½-5 items and 25 items in the C-TRF (see sections below for information regarding testing time and internal consistency). Items that did not fit together on any scale have been retained as "Other Problems" across the revisions of the measures.

II. Administration of Measure

Who is the Respondent to the Measure?

Parent. The CBCL/1½-5 is designed to be completed by parents or others who see children regularly in a home setting.

Teacher or caregiver. The C-TRF is designed to be completed by individuals who observe and interact with the child on a regular basis in a preschool or daycare setting with at least 3 other children, and who have known the child for a minimum of 2 months.

Achenbach and Rescorla (2000) indicate that respondents to ASEBA measures should have at least fifth grade reading skills.

If Child is Respondent, What is Child Asked to Do?
Not applicable.

Who Administers Measure/Training Required?
Test Administration:

- Because these measures are usually administered as written questionnaires, little specific training is required for actual administration.

Data Interpretation:

- Achenbach & Rescorla (2000, 2001) suggest that graduate training at the Master's level or higher, or two years of residency in pediatrics, psychiatry, or family medicine are usually needed for interpretation of results.

Setting (e.g., one-on-one, group, etc.)
One-on-one or independently. These measures are typically administered as questionnaires that parents or teachers complete on their own. Alternatively, if a respondent has reading difficulties the measures can be administered by an interviewer who reads the measure aloud to the respondent and records the respondent's answers. In fact, Achenbach and Rescorla (2000) indicate that standard administration of the CBCL/1½-5 with the normative sample involved reading the measure aloud to parents.

Time Needed and Cost
Time:

- Both the CBCL/1½-5 and the C-TRF take approximately 10-15 minutes to complete.

Cost:

- Manuals for the CBCL/1½-5 and the C-TRF combined, and for the CBCL/6-18 and the TRF combined: $35.00 each
- Hand-scored forms: $25 for packages of 50
- Reusable templates for hand-scoring: $7 each
- Scannable forms available for the CBCL/6-18 and the TRF: $45 for 50 forms
- Scoring software ranges from $170 to $250 for a single-user license. Scanning software and options for direct client computer entry and for on-line administration are also available for the CBCL/6-18 and the TRF.

Comments

- The fifth grade reading level that is indicated for these measures may present some problems for very low SES, high-risk samples. However, Achenbach and Rescorla

(2000) provide specific instructions for administering the questionnaire in an interview format that minimizes possible embarrassment and that they also suggest will minimize error due to nonstandard administration.

III. Functioning of Measure

Reliability Information from Manual

Internal consistency

- Internal consistency statistics for the CBCL/1½-5 syndromes and scales were calculated for a sample of 563 children who had been referred to 14 mental health and special education programs and an equal number of children from the normative sample who were selected to match the referred children as closely as possible with respect to age, gender, SES, and ethnicity. Cronbach's alphas for syndromes ranged from .66 (Anxious/Depressed) to .92 (Aggressive Behavior). Alphas for the DSM-Oriented scales ranged from .63 (Anxiety Problems) to .86 (Oppositional Defiant Problems). The alpha for the Internalizing scale was .89, alpha for the Externalizing scale was .92, and alpha for the Total Problems scale was .95 (see Achenbach & Rescorla, 2000, p. 155-156).
- Information regarding the internal consistency of the C-TRF syndromes and scales was reported based on a sample including 303 children who had been referred to 11 mental health and special education programs and 303 matched children from the normative sample. Coefficient alphas for the C-TRF syndromes ranged from .52 (Somatic Complaints) to .96 (Aggressive Behavior). Alphas for the DSM-Oriented scales ranged from .68 (Anxiety Problems) to .93 (Oppositional Defiant Problems). The Internalizing scale had an alpha of .89, alpha for the Externalizing scale was .96, and the Total Problems scale had an alpha of .97 (see Achenbach & Rescorla, 2000, p. 157-158).

Cross-informant agreement

- Agreement between mother- and father-report on the CBCL/1½-5 was examined in a sample of 72 children, some of whom had been referred to clinical services. Mean scale scores of mothers and fathers were not significantly different. Correlations between maternal and paternal ratings ranged from .48 to .66 for the syndromes, and from .51 to .67 for the DSM-Oriented scales. Inter-parent correlations were .59 for Internalizing, .67 for Externalizing, and .65 for Total Problems. The mean correlation was .61 (see Achenbach & Rescorla, 2000, p. 78).
- Agreement between different caregiver or teachers on the C-TRF was computed in a sample of 102 children, including participants in the NICHD Study of Early Child Care and other children attending preschools in Vermont and The Netherlands. With one exception, correlations ranged from .52 to .78 for the syndromes and were similar to those found between mothers and fathers; the cross-teacher correlation for Somatic Complains syndrome was .21. Correlations ranged from .55 to .71 for the DSM-Oriented scales. Internalizing was correlated .64 across teachers, Externalizing was correlated .79, and Total Problems was correlated .72. The mean correlation was .65 (see Achenbach & Rescorla, 2000, p. 78).
- Agreement between parents and caregivers or teachers was computed for a sample of 226, some included in the 1999 National Survey and others obtained from clinical settings.

Interrater correlations ranged from .28 to .55 for the syndromes, and from .21 to .52 for the DSM-Oriented scales. Parent and teacher ratings on Internalizing were correlated .30. There was a .58 correlation for Externalizing, and the Total Problems correlation was .50. The mean correlation was .40 (see Achenbach & Rescorla, 2000, p. 78).

Test-retest reliability

- Test-retest reliabilities of the CBCL/1½-5 syndromes and scales were examined in a sample of 68 nonreferred children from 3 U.S. sites whose mothers completed the CBCL/1½-5 twice, approximately 8 days apart. Correlations across the two ratings ranged from .68 (Anxious/Depressed) to .92 (Sleep Problems) for the syndromes, and from .74 (Attention Deficit/Hyperactivity Problems) to .87 (Oppositional Defiant Problems) for the DSM-Oriented scales. The test-retest correlations for Internalizing, Externalizing, and Total Problems scales were .90, .87, and .90, respectively. The mean test-retest correlation was .85 (see Achenbach & Rescorla, 2000, p. 76).
- For the C-TRF, test-retest reliabilities were estimated for a sample of 59 nonreferred children who were rated by their preschool caregivers. Again, the testing interval was approximately 8 days. Of this sample, 39 were in The Netherlands, while the remaining 20 children attended a preschool in Vermont. Test-retest correlations ranged from .68 (Anxious/Depressed) to .91 (Somatic Complaints) for the syndromes, and from .57 (Anxiety Problems) to .87 (Oppositional Defiant Problems) for the DSM-Oriented scales. Correlations were .77 for Internalizing, .89 for Externalizing, and .88 for Total Problems. The mean test-retest correlation was .81 (see Achenbach & Rescorla, 2000, p. 76).
    - Longer-term stability of the CBCL/1½-5 was examined in a sample of 80 children whose mothers rated their children a second time 12 months after initially completing the CBCL/1½-5. Correlations for the syndromes ranged from .53 (Withdrawn) to .64 (Anxious/Depressed), correlations for the DSM-Oriented scales ranged from .52 (Pervasive Developmental Problems) to .60 (Anxiety Problems), and the correlations for Internalizing, Externalizing, and Total Problems were .76, .66, and .76, respectively. The mean correlation was .61 (see Achenbach & Rescorla, 2000, p. 80).
    - Finally, stability of C-TRF ratings across 3 months was examined in a small sample of 32 preschoolers enrolled in one preschool program. In this very small sample of children attending a single preschool, cross-time correlations varied considerably across the scales. Correlations for the syndromes ranged from .22 (nonsignificant; Somatic Complaints) to .71 (Emotionally Reactive). Correlations for the DSM-Oriented scales ranged from .46 (Attention Deficit/Hyperactivity Problems) to .85 (Affective Problems), and correlations were .65, .40, and .56 for Internalizing, Externalizing, and Total Problems, respectively. The mean correlation was .59 (see Achenbach & Rescorla, 2000, p. 80).

Validity Information from Manual

Convergent validity

- Achenbach and Rescorla (2000) reported correlations in their own work and in independent studies (Spiker, Kraemer, Constantine, & Bryant, 1992; Koot, van den Oord, Verhulst, & Boomsma, 1997) ranging from .56 to .77 between an earlier preschool version of the

CBCL—the CBCL/2-3—and the Behavior Checklist (BCL; Richman, Stevenson, & Graham, 1982; see Achenbach & Rescorla, 2000, p. 97).

- Additional studies examined convergence between earlier versions of the CBCL and other measures. Mouton-Simien, McCain, & Kelly (1997) found a correlation of .70 between CBCL Total Problems scale scores and a sum of frequency ratings on the Toddler Behavior Screening Inventory. Briggs-Gowan and Carter (1998) reported correlations ranging from .46 to .72 between externalizing scale scores from the CBCL and their Infant-Toddler Social and Emotional Assessment, and correlations between internalizing scales from the two measures ranging from .48 to .62 (see Achenbach & Rescorla, 2000, p. 97).
- Two studies indicated moderate significant correlations between preschoolers' CBCL scale scores and DSM diagnostic work. Keenan and Wakslag (2000) found a correlation of .49 between CBCL Externalizing scores and a summary score of DSM Opppositional Defiant Disorder and Conduct Disorder symptoms assessed during diagnostic interviews with mothers. Arend, Lavigne, Rosenbaum, Binns, and Christoffel (1996) found a correlation of .47 between the CBCL/2-3 Aggressive Behavior scale and DSM diagnoses of disruptive disorders (see Achenbach & Rescorla, 2000, p. 97).

Discriminant validity

- Achenbach and Rescorla (2000) discussed two studies (Achenbach, Edelbrock, & Howell, 1987; Koot et al., 1997) in which CBCL/2-3 scores were correlated with developmental measures, including the Bayley (1969) Mental Scale, the McCarthy (1972) General Cognitive Index, and the Minnesota Child Development Inventory. These measures were designed as assessments of development, while scores on the CBCL/2-3 were expected to be to some extent independent of developmental level. In neither study were these measures significantly correlated with the CBCL/2-3 (see Achenbach & Rescorla, 2000, p. 99).

Predictive validity

- CBCL/2-3 data from an earlier study (Achenbach, Howell, Aoki, & Rauh, 1993) was rescored according to new CBCL/1½-5 guidelines, to look at the extent to which preschoolers' scores at ages 2 and 3 predicted scores obtained yearly from ages 4 to 9 on the CBCL/4-18 (since revised and renormed as the CBCL/6-18). Across-time correlations were all significant for Internalizing Behavior, Aggressive Behavior, Externalizing, and Total Problems, ranging from .39 (for Internalizing Behavior at ages 2 and 5) to .75 (for Total Problems at ages 3 and 4). Correlations were not consistently significant for Anxious/Depressed, Somatic Problems, Withdrawn, and Attention Problems, particularly when predicting later scores from age 2 assessments. Correlations for age 2 assessments ranged from .05 (for Attention Problems at ages 2 and 7) to .51 (for Attention Problems at ages 2 and 4). Correlations for age 3 assessments ranged from .10 (for Somatic Problems at ages 3 and 4) to .56 (for Attention Problems at ages 3 and 4; see Achenbach & Rescorla, 2000, p. 98).

Criterion validity

- All of the CBCL/1½-5 raw scale scores were significantly associated with referral status in a sample of 563 referred children and 563 nonreferred children (described earlier in the section on internal consistencies). In all cases, the referred children had higher syndrome and scale

scores than the nonreferred children. The manual presents associations between referral status and CBCL/1½-5 syndrome and scale scores in terms of the percent of variance accounted for (r^2) in scale scores by referral status. The strongest associations were between referral status and Pervasive Developmental Problems ($r^2 = .25$), Total Problems ($r^2 = .22$), Affective Problems ($r^2 = .20$), and Internalizing ($r^2 = .20$). The weakest associations (r^2 lower than .10) between referral status and CBCL/1½-5 raw scores were with Anxious Depressed, Sleep Problems, Attention Problems, Aggressive Behavior, Externalizing, Anxiety Problems, Attention Deficit/ Hyperactivity Problems, and Oppositional Defiant Problems (see Achenbach & Rescorla, 2000, p. 85).

- Similarly, all of the C-TRF raw scores were significantly associated with referral status in a sample including 303 referred children and an equal number of nonreferred children (see section on internal consistencies for further description of this sample). The strongest associations between referral status and C-TRF scores were for Total Problems ($r^2 = .24$), Externalizing ($r^2 = .23$), Aggressive Behavior ($r^2 = .22$), and Oppositional Defiant Problems ($r^2 = .20$), while the weakest associations were with Anxious/Depressed, Somatic Complaints, Withdrawn, Affective Problems, and Anxiety Problems (see Achenbach & Rescorla, 2000, p. 85). Additional analyses using odds ratios supported these findings. Referred children were more likely to have CBCL/1½-5 and C-TRF scores in the clinical range than were nonreferred children. Odds ratios were significant for all scales (see Achenbach & Rescorla, 2000, p. 91).

Reliability/Validity Information from Other Studies

There are large numbers of studies within the developmental and child clinical research literatures that have used ASEBA measures, including studies of preschoolers, using either the CBCL/2-3 for younger children or the CBCL/4-18 and TRF/4-18 for older preschoolers. Because the newly revised measures are not dramatically different from older versions, and because raw scores that are not adjusted according to age norms are typically used in research, these studies can be used to examine the validity of current ASEBA measures, as well as their usefulness for research and applied purposes.

Two recent studies included information on associations between CBCL scales and scales from the Social Skills Rating System (SSRS; Gresham & Elliot, 1990):

- Using a sample of children enrolled in Head Start programs, Kaiser, Hancock, Cai, Foster, and Hester (2000) reported significant correlations ranging from .54 to .65 between the CBCL/2-3 Internalizing, Externalizing, and Total Problem Behavior scales and parallel scales from the SSRS.
- In another study, Gilliom, Shaw, Beck, Schonberg, and Lukon (2002) reported a significant negative correlation between CBCL/6-18 Externalizing behavior problems and Cooperation as assessed with the SSRS at age 6 in a sample of boys from low income families. High levels of Externalizing at age 6 were also predicted by observations of poor anger regulation at age 3½.

Comments

- Cronbach's alpha coefficients indicate high internal consistency for the three summary scales from both the CBCL/1½-5 and the C-TRF. Alphas were considerably lower for some of the

syndromes and DSM-IIII-Oriented scales. Among the syndromes and scales, those tapping externalizing problems (particularly Aggressive Behavior and Oppositional Defiant Problems) generally had higher alphas than did syndromes and scales tapping internalizing problems (particularly Anxious/Depressed and Anxiety Problems). The Somatic Complaints syndrome had low internal consistency (.52) on the C-TRF, but was more internally consistent (.80) on the parent-report CBCL/1½-5.

- With respect to cross-informant agreement, correlations between mother- and father-ratings were all moderate to high, and all but one correlation (Somatic Complaints) between ratings by different teachers were also high. As might be expected given the very different nature of their interactions with the children being evaluated, correlations between parent CBCL/1½-5 ratings and caregiver or teacher C-TRF ratings were lower than either inter-parent or inter-caregiver correlations.
- Test-retest correlations indicate strong consistency in both teacher- and parent-reports of children's behavior problems across an 8-day interval. As would be expected, cross-time consistency in children's relative scores was somewhat lower across a 12-month interval for the CBCL/1½-5, and across a 3-month interval for the C-TRF. It is interesting to note that the Somatic Complaints syndrome appeared to be the least reliable measure from the C-TRF, as reflected in all but one of the reliability indicators (i.e., internal consistency, cross-informant agreement, 3-month test-retest correlations); however, Somatic Complaints actually had the highest 8-day test-retest correlation on the C-TRF.
- Taken together, information on convergent, discriminant, and criterion validity provides support for the use of the CBCL/1½-5 and the C-TRF as measures of behavior problems in the preschool period.
- No information was found regarding the effects on scores, reliability, or validity of the ASEBA instruments when administered as interviews, compared with the standard written format. This information may be of substantial importance, particularly when the CBCL/1½-5 is used with samples of children from low income, poorly educated families.
    - ASEBA measures have been criticized in the past for having high correlations between Internalizing and Externalizing scales. Reported correlations are .22 and .59 for CBCL/1½ -5 correlations between Internalizing and Externalizing for referred and non-referred children, respectively, and correlations between the two scales on the C-TRF are .53 and .62, suggesting that the two scales do not tap distinctly different types of behavioral and emotional problems, particularly within a general population of preschoolers (see Achenbach & Rescorla, 2000, pp. 159-160). These correlations are between standardized (*T*) scores, and may be lower than correlations between raw scores, which are not reported. Two other measures from which internalizing and externalizing scores are derived—the Social Competence and Behavior Evaluation (SCBE; LaFreiniere & Dumas, 1995) and the Behavioral Assessment System for Children (BASC; Reynolds & Kamphaus, 1992)—both report lower correlations between internalizing and externalizing scores compared with those reported for ASEBA measures.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- Behavioral and emotional problems, as tapped by ASEBA measures, have been found to be more prevalent among children living in poverty than among children living in higher income families (e.g. Briggs-Gowan, Carter, Moye Skuban, & McCue Horwitz, S., 2001).
- Shaw, Keenan, Vondra, Delliquadri, and Giovannelli (1997) reported associations between CBCL Internalizing scores and a set of child and family risk factors assessed in infancy in a sample of 86 low-income children and their mothers. Internalizing scores were predicted by infant negative emotionality (temperament), by disorganized infant-mother attachment classification, and by mother-reported negative life events, childrearing disagreements, and parenting hassles. Children who were temperamentally prone to negative emotionality were particularly negatively affected by exposure to parental conflict.

V. Adaptations of Measure

The Behavior Problems Index (BPI)

Description of Adaptation

The BPI was developed by N. Zill and J. Peterson as a brief (28 item) measure of behavioral adjustment appropriate for use in large-scale surveys (see Peterson & Zill, 1986). Items were adapted from the CBCL as well as other behavior problems measures. The BPI was originally designed for and included in the 1981 Child Health Supplement of the National Health Interview Survey (NCHS, 1982). A description of the BPI and its use in the NLSY – Child is available from the Center for Human Resource Research, The Ohio State University (e.g. Baker, Keck, Mott, & Quinlan, 1993).

Originally developed as a parent report, a parallel teacher report version was included in the New Chance Evaluation study. The BPI can be used with children 4 years of age and older.

One total Behavior Problems Index is derived from the BPI. There are also six behavioral subscales, identified based on factor analyses:

- Antisocial.
- Anxious/Depressed.
- Headstrong.
- Hyperactive.
- Immature/Dependency.
- Peer Conflict/Social Withdrawal.

In addition, some studies use Externalizing and Internalizing Behavior Problems subscales instead of the six originally identified (e.g. Gennetian & Miller, 2002).

Psychometrics of Adaptation

Single-year age norms were developed from the 1981 National Health Interview Survey administration for all children and for males and females separately. These norms are based on

binary data (although mothers responded on a 3-point scale), and subscales each have relatively few items. Thus, there is a limited range of both raw and normed scores. Norms tables are available in the NLSY Child Handbook, Revised Edition (Baker, et al., 1993). The majority of studies that utilize the BPI, however, do not appear to use normed data, but rather utilize raw data, either scored with the full 3-point item response scales or converted to 2-point scales indicating presence or absence of behaviors described.

Reliability

- In the 1981 NHIS sample, internal consistency (Cronbach's alpha) of the Total BPI score was .89 for children (ages 4-11) and .91 for adolescents ages 12-17).
- In the 1990 NLSY – Child survey, alpha for the Total BPI score was .88. Subscale alphas ranged from .57 (for Peer Conflict, a 3-item scale) to .71 (for Headstrong, a 5-item scale). Similar reliability estimates were found for the 1986 and 1988 NLSY – Child samples.

Validity

Findings from the NLSY – Child sample:

- Correlations between BPI subscales range from .29 to .58 in the 1990 survey (median = .42), indicating that the subscales tap relatively distinctive problem behavior components (Baker et al., 1993).
- Correlations between BPI Total and subscale scores across 2 years (1986 to 1988) ranged from .33 to .59. Across 4 years (1986 to 1990), correlations ranged from .32 to .49. In both cases, the highest correlations were for the BPI Total score. These correlations compare favorably to scale and subscale scores from other behavior problems measures (Baker et al., 1993)
- High (negative) scores on the BPI were associated with low levels of observed cognitive stimulation and emotional support in the home. Correlations were significant but fairly low (-.22 and -.17 with Cognitive Stimulation and Emotional Support as assessed with the adapted version of the HOME instrument used in the NLSY; Baker et al., 1993).
- A number of researchers working independently with NLSY – Child data have reported significant relationships between BPI scores and social and demographic variables in this sample, including low family income and poverty status (e.g. Dubow & Luster, 1990; Vandell & Ramanan, 1991).

In a recent study, Gennetian and Miller (2002) examined outcomes for children ages 5 to 13 whose mothers had been randomly assigned three years earlier to participate in an experimental welfare program (the Minnesota Family Investment Program; MFIP) that provided financial incentives to encourage work (e.g., more benefits retained than under normal AFDC guidelines), coupled with mandatory participation in employment-related activities, compared with children whose mothers had been assigned to receive traditional AFDC benefits. Among their findings was that children whose mothers participated in MFIP were rated by their mothers as having significantly fewer BPI Externalizing Problems than were children of mothers receiving traditional AFDC benefits. This impact was more pronounced for children who were 6 years of age or older at the time of random assignment (9 to 13 years of age at time of assessment) than for younger children.

`

Studies Using Adaptation

- National Health Interview Survey, 1981 Child Health Supplement.
- National Survey of Children, 1981.
- National Longitudinal Survey of Youth, Children of the NLSY, 1986 and subsequent.
- New Chance Evaluation.
- Child Outcomes Study of the National Evaluation of Welfare-to-Work Strategies (Two Year Follow-up).

Early Childhood Measures: Social-Emotional

Conners' Rating Scales–Revised (CRS-R)

I. Background Information

Author/Source

Source: Conners, C. K. (1997). *Conners' Rating Scales – Revised: Technical manual.* North Tonawanda, NY: Multi-Health Systems, Inc.

Publisher: Multi-Health Systems, Inc.
P.O. Box 950
North Tonawanda, NY 14120-0950
Phone: 800-456-3003
Website: www.mhs.com

Purpose of Measure

As described by instrument publisher:
The CRS-R is an assessment of behavior and emotional disorders of childhood, particularly Attention Deficit Hyperactivity Disorder (ADHD). The CRS-R can be used "…as a screening measure, treatment monitoring device, research instrument, and direct clinical/diagnostic aid" (Conners, 1997, p. 5).

Population Measure Developed With

- The norming sample for the long form of the parent scale (CPRS-R:L) included 2,482 children and adolescents between the ages of 3 and 17 with approximately equal numbers of males and females. Approximately 15 percent were ages 3 to 5, 26 percent were 6 to 8 years old, 22 percent were 9 to 11 years old, 22 percent were ages 12 to 14, and 15 percent were 15 to 17 years old. Each child was rated by a parent or guardian. Ethnic information for parents indicates that 83 percent were white, 4.8 percent were black, 3.5 percent were Hispanic, 2.2 percent were Asian/Pacific Islander, 1.1 percent were American Indian/Alaskan Native, and 4.9 percent indicated another ethnicity or did not provide any ethnicity information. No direct information on children's race/ethnicity was reported. The median annual income of the participating families was between $40,001 and $50,000. No information was provided by Conners (1997) on mother's or father's education.
- A total of 2,426 cases were used to develop norms for the short form of the parent scale (CPRS-R:S), the majority of which were drawn from the CPRS-R:L norming sample. Race/ethnicity and gender characteristics of the two samples were very similar. Approximately 12 percent of the children were ages 3 to 5, 26 percent were 6 to 8 years old, 23 percent were 9 to11 years old, 23 percent were ages 12 to 14, and 16 percent were 15 to 17 years old.
- The norming sample for the long form of the teacher scale (CTRS-R:L) included 1,973 children and adolescents between the ages of 3 and 17 (49 percent male, 51 percent female), each rated by one of their teachers. No information was provided on the total number of teachers who performed ratings. Teachers identified 78 percent of the children

as white, 10.2 percent as black, 5.8 percent as Hispanic, 1.6 percent as Asian/Pacific Islander, 1.5 percent as American Indian/Alaskan Native, and 2.8 percent as other or no ethnic background information provided. No other family demographic data were provided by Conners (1997). Of the 1,973 children, 10 percent were ages 3 to 5, 27 percent were 6 to 8 years old, 25 percent were 9 to 11 years old, 26 percent were ages 12 to 14, and 12 percent were 15 to 17 years old.

- As with the parent rating samples, the majority of the 1,897 children included in the norming sample for the short form of the teacher scale (CTRS-R:S) were drawn from the long form norming sample. A smaller percentage of the children in this sample than in the long form norming sample were identified as black (7.2 percent) and a larger percentage (81 percent) were white. Approximately 6 percent of the children were ages 3 to 5, 29 percent were 6 to 8 years old, 26 percent were 9 to 11 years old, 27 percent were ages 12 to 14, and 13 percent were 15 to 17 years old.

Age Range Intended For

Children and adolescents ages 3 years through 17 years.

Key Constructs of Measure

There are long and short forms of the CRS-R for both parents (the CPRS-R:L and the CPRS-R:S) and teachers (the CTRS-R:L and the CTRS-R:S). All of the subscales of the CRS-R pertain to behavioral and emotional problems. There are no dimensions of positive functioning assessed with the CRS-R. The long forms include six or seven factor analytically derived subscales, as well as a series of "Auxiliary Scales":

- *Factor analytically derived subscales*
  - *Oppositional:* Rule-breaking, noncompliance, and tendencies toward annoyance and anger.
  - *Cognitive Problems/Inattention:* Inattentiveness, poor concentration, difficulties completing school-related tasks and other activities, and poor academic performance.
  - *Hyperactivity:* Restlessness, excitability, and impulsivity.
  - *Anxious-Shy:* Fearfulness, timidity, shyness, and sensitivity to slights and criticism.
  - *Perfectionism*: Excessive focus on details, fastidiousness, and inability to adapt to change.
  - *Social Problems:* Poor social skills, inability to make and keep friends.
  - *Psychosomatic*: This scale is included in the CPRS-R:L only. Aches and pains, general illness and fatigue.
- *Auxiliary scales*
  - *Conners' Global Index (CGI):* A modified version of the Hyperactivity Index from the original CRS. Originally a single scale, factor analytic studies have indicated that the CGI includes two separate components. The items on this scale do not overlap with items on factor analytically derived subscales.
    - *Restless-Impulsive:* Restlessness, excitability, disruptiveness, and attention problems.
    - *Emotional Lability:* Moodiness, frequent and intense expressions of negative emotion.

- *ADHD Index*: A set of 12 items that were specifically selected because of their combined ability to identify children with a diagnosis of ADHD, and thus, to serve as a screening instrument for ADHD. Items primarily involve restlessness and problems with attention and distractibility. Some of the items on this scale are also found on the factor analytically derived subscales.
- *DSM-IV Symptoms subscales*: The items on this scale, and the two subscales of which it is composed, directly relate to the 18 criteria for clinical diagnosis of ADHD, hyperactive-impulsive or inattentive types, as defined in the fourth edition of the Diagnostic and Statistical Manual for Mental Disorders (DSM-IV; American Psychiatric Association, 1994).
  - *DSM-IV Inattentive:* Forgetfulness, distractibility, poor organizational abilities.
  - *DSM-IV Hyperactive-Impulsive:* Excessive movement and talking, poor impulse control.

The CRS-R short forms include abbreviated numbers of items for four of the scales listed above: Oppositional, Cognitive Problems/Inattention, Hyperactivity, and the ADHD Index. In addition, the Conners' Global Index, the ADHD Index, and the DSM-IV Symptoms subscales can be administered independently.

Norming of Measure (Criterion or Norm Referenced)

Norm referenced. There are separate norms for males and females in each of five age groups: 3 to 5 years, 6 to 8 years, 9 to 11 years, 12 to 14 years, and 15 to 17 years.

The DSM-IV symptoms subscales can also be used as a criterion-referenced diagnostic tool. Children for whom at least six of the nine behaviors on one of the subscales (DSM-IV Inattentive or DSM-IV Hyperactive-Impulsive) are rated as being "Very Much True (Very Often, Very Frequent)," the highest rating on the CRS-R, may meet DSM-IV criteria for the associated type of ADHD (predominantly inattentive or predominantly hyperactive-impulsive). Children who meet criteria for both subtypes may meet criteria for a combined-type ADHD diagnosis.

Comments

- This measure heavily emphasizes diagnosis of ADHD and recognition of subclinical attention and hyperactivity problems, and there are no positive behaviors assessed with either the short or long forms.
- Information on family demographic characteristics other than ethnicity were not consistently provided for all of the norming samples. It does appear, however, that the samples were relatively affluent, and that ethnic minorities were substantially underrepresented.
- Although there are separate norms for children of different ages, the subscales are the same at all ages. While this can be a positive feature, allowing for clearer comparisons across ages, it is not clear to what extent the items were evaluated for relevance at all ages covered by the CRS-R.

II. Administration of Measure

Who is the Respondent to the Measure?
Parent, teacher and, adolescent. There are separate parent report and teacher report forms of the CRS-R as well as an adolescent self-report form.

If Child is Respondent, What is Child Asked to Do?
N/A.

Who Administers Measure/ Training Required?
Test Administration:

- The CRS-R is generally administered as a questionnaire, and no specialized training is required. Individuals responsible for administration and interpretation should "…be members of professional associations that endorse a set of standards for the ethical use of psychological or educational tests or licensed professionals in the areas of psychology, education, medicine, social work, or an allied field" (Conners, 1997, p. 8).
- Conners (1997) indicates that teachers must have adequate time to become knowledgeable about a child prior to rating and recommends that administration should not occur until at least 1 to 2 months after the beginning of the school year.
- The reading level of both parent and teacher versions of the CRS-R (long and short forms) is relatively high. Readability analyses indicate that items are written at a 9th or 10th grade reading level.

Data Interpretation:

- According to the manual, "interpretation must be assumed by a mature professional who realizes the limitations of such screening and testing procedures" (Conners, 1997, p. 8).

Setting (e.g., one-on-one, group, etc.)
Parents and teachers usually complete rating scales on their own, preferably in a single sitting. Group administration is possible as well.

Time Needed and Cost
Time:

- Administration of the CTRS-R:L takes approximately 15 minutes (per student), and the CPRS-R:L generally takes 15-20 minutes. Short versions of the scales can be completed in 5-10 minutes.

Cost:

- Manual (including information for all forms): $46.00
- Parent and Teacher long forms: $29.00 per package of 25
- Parent and Teacher short forms: $27.00 per package of 25

Comments

- The reading level required is of concern, particularly when parents with low education levels are asked to be respondents. Even if read to a low-level reader, some items may not be well understood.

III. Functioning of Measure

Reliability Information from Manual

Internal Consistency

Conners (1997) presents internal consistency coefficients (Cronbach's alphas) for males and females separately by age group. In the following summary, we will focus on reliability estimates for the two youngest groups (3 to 5 and 6 to 8; see Connors, 1997, p.113-114).

- *CPRS-R:L.* Internal reliabilities for both factor analytically derived and auxiliary scales were similar for males and females, and for younger and older children. Alphas fell below .80 for only two scales—Psychosomatic and CGI Emotional Lability.
    - Three to Five Year Olds
        - Within the factor analytically derived scales, coefficient alphas ranged from .76 to .90 (median = .84) for males and .75 to .88 (median = .86) for females.
        - Within the Auxiliary scales coefficient alphas ranged from .69 to .94 (median = .89) for males and .77 to .91 (median = .86) for females.
    - Six to Eight Year Olds
        - Within the factor analytically derived scales, coefficient alphas ranged from .75 to .93 (median = .85) for males and .81 to .93 (median = .91) for females.
        - Within the Auxiliary scales coefficient alphas ranged from .80 to .95 (median = .89) for males and .76 to .94 (median = .92) for females.
- *CPRS-R:S.* Coefficient alphas for the three factor analytically derived subscales (Oppositional, Cognitive Problems/Inattention, and Hyperactivity) and the ADHD Index included in the short form of the parent report ranged from .87 to .93 for younger males, from .83 to .89 for younger females, from .88 to .94 for older males, and from .88 to .94 for older females.
- *CTRS-R:L.* As can be seen from the information presented below, alphas were generally similar for teacher report scales as for those reported for the CPRS-R:L. However, in the younger age group, alphas were consistently lower for ratings of girls than for boys, with the exception of ratings on the Anxious-Shy scale.
    - Three to Five Year Olds
        - Within the factor analytically derived scales, coefficient alphas ranged from .84 to .95 (median = .87) for males and .59 to .83 (median = .80) for females.
        - Within the Auxiliary scales coefficient alphas ranged from .78 to .96 (median = .93) for males and .74 to .87 (median = .82) for females.
    - Six to Eight Year Olds

  - Within the factor analytically derived scales, coefficient alphas ranged from .82 to .94 (median = .91) for males and .84 to .93 (median = .91) for females.
  - Within the Auxiliary scales coefficient alphas ranged from .79 to .96 (median = .95) for males and .77 to .96 (median = .93) for females.

- *CTRS-R:S.* Alphas for the three factor analytically derived scales and the ADHD Index from the teacher report, short form, ranged from .85 to .97 for younger males, from .81 to .86 for younger females, from .87 to .96 for older males, and from .89 to .94 for older females. Consistent with findings for the long form, alphas for teacher ratings of younger girls were lower than alphas for ratings of younger males.

Test-Retest Reliability

Test-retest reliabilities for both parent and teacher ratings were conducted with small samples of children and adolescents (49 parent ratings, 50 teacher ratings). No separate analyses were reported for different age groups or for males and females. Ratings were conducted 6 to 8 weeks apart. The same samples were used for short and long forms. Rather than completing separate short forms, short form subscales were derived from the long form versions of the measures. Overall, test-retest correlations were moderate to high across this fairly short interval (see Connors, 1997, p.113-114).

- *CPRS-R:L.* Test-retest correlations of the 14 scales ranged from .47 (Anxious-Shy) to .85 (Hyperactivity).
- *CPRS-R:S.* Test-retest correlations for the four short form scales ranged from .62 (Oppositional) to .85 (Hyperactivity).
- *CTRS-R:L.* Test-retest correlations for the 13 scales ranged from .47 (both Cognitive Problems/Inattention and DSM-IV Hyperactive-Impulsive) to .88 (Anxious-Shy).
- *CTRS-R:S.* Test-retest correlations for the four short form subscales ranged from .72 (Hyperactivity) to .92 (Cognitive Problems/Inattention).

Interrater Reliability

A subsample of 501 male and 523 female children and adolescents from the norming sample were rated by both a parent and a teacher. Correlations between parallel parent- and teacher-report subscales varied widely (see Connors, 1997, p.128-129).

- For the long forms, parent-teacher correlations between the six parallel factor analytically derived subscales ranged from .12 to .47 for males and from .21 to .55 for females. For both males and females, the highest levels of agreement were for Cognitive Problems/Inattention, while the lowest agreement was found for Perfectionism. Correlations among the auxiliary scales ranged from .28 (CGI Emotional Lability) to .50 (CGI Restless-Impulsive) for males, and from .16 (CGI Emotional Lability) to .49 (ADHD Index) for females.
- Parent-teacher correlations between the three parallel factor analytically derived subscales on the short forms ranged from .33 to .49 for males, and from .18 to .52 for females. For both males and females, the highest levels of agreement were again found for Cognitive Problems/Inattention, while the lowest levels of agreement were found for Oppositional. For both males and females, the interrater correlation for the ADHD Index was .49.

Validity Information from Manual

Factorial Validity

The two long forms of the CRS-R include scales constructed on the basis of factor analyses of responses from participants in pilot studies. Modifications in the forms were undertaken and additional exploratory factor analyses of data from larger independent samples were undertaken, followed by confirmatory factor analyses with additional independent samples. According to Connors (1997), the goal of the factor analytic procedures was to develop subscales representing "…distinct dimensions of problem behavior and psychopathology" (p. 121). The factorial validity (a form of construct validity) of the resulting subscales was addressed in the norming sample by examining the intercorrelations among the factor-derived subscales. The remaining auxiliary scales were not designed to be independent, but were rather conceptualized as different approaches to assessing ADHD.

- The CPRS-R:L includes seven factor analytically derived subscales. Correlations between these subscales were conducted separately for males and females. For males, correlations ranged from -.01 to .59, with a median correlation of .36. For females, correlations among the seven subscales ranged from -.02 to .52, with a median correlation of .35 (see Connors, 1997, p. 122). The lowest correlations were between Perfectionism and all other subscales (ranging from -.01 to .26 for males, from -.02 to .24 for females). The highest correlations were between Cognitive Problems, Oppositional, and Hyperactivity (ranging from .51 to .59 for males, from .49 to .52 for females).
- The CTRS-R:L includes six factor analytically derived scales. For males, correlations among the six scales ranged from -.08 to .63, with a median correlation of .39. For females, correlations ranged from -.15 to .54, with a median correlation of .26 (see Connors, 1997, p. 124).
- The factorial validities of the three factor analytically derived scales included in both of the short forms of the CRS-R (Oppositional, Cognitive Problems/Inattention, and Hyperactivity) were examined using confirmatory factor analytic procedures. Goodness of fit indices for both the CPRS-R:S and the CTRS-R:S indicated adequate fits of the data to the three-factor models (see Connors, 1997, pp. 122-123 and 124-125).
    - For the CPRS-R:S, correlations between the three factors ranged from .53 to .56 for males, and from .48 to .49 for females (see Connors, 1997, p. 123).
    - Correlations between the three CTRS-R:S factors ranged from .38 to .63 for males, and from .31 to .55 for females (see Connors, 1997, p. 125).

Convergent Validity

Conners (1997) reported results from two small samples of children who were asked to complete the Children's Depression Inventory (CDI; Kovacs, 1992). In one sample of 33 children and adolescents with a mean age of 10.39 (*SD* = 2.46), parents were asked to complete the CPRS-R:L. In the second sample of 27 children and adolescents with a mean age of 10.41 (*SD* = 2.47), teachers were asked to complete the CTRS-R:L. Although the full age ranges of children in these studies were not specified, the CDI cannot be administered to very young children (see Connors, 1997, p.133). Associations between the CDI and the CPRS-R:L and the CTRS-R:L were examined "…to check for positive associations between the various Hyperactivity subscales on the CRS-R and negative dysphoria. Such associations would be consistent with

well-established descriptions in the developmental literature of the hyperactive-impulsive-emotionally labile child…" (Conners, 1997, p. 132).

- Results of the study examining associations between parental ratings on the CPRS-R:L and children's self-reports on the CDI indicated that, with the exception of Perfectionism, all of the CPRS-R:L subscales had statistically significant correlations with CDI Total scores (ranging from .38 for Anxious-Shy to .82 for Oppositional), and generally showed correlations of similar strength with 3 or more of the 5 CDI subscales. Of the CDI subscales, Negative Mood, Ineffectiveness, and Anhedonia were consistently correlated significantly with the remaining CPRS-R:L scales, while fewer significant correlations existed between CPRS-R:L scales and the CDI subscales Interpersonal Problems and Negative Self-Esteem. SPRS-R:L Perfectionism was not significantly correlated with CDI Total Scores (r = .23), or with any of the CDI subscales.
- Associations between teacher ratings on the CTRS-R:L and children's CDI self-reports were similar to those between the CPRS-R:L and the CDI. Of the 13 CTRS-R:L scales, 10 were significantly correlated with the CDI Total score (correlations ranging from .41 to .69). Perfectionism was again uncorrelated with the CDI Total score or with any CDI subscale. The DSM-IV Hyperactive-Impulsive scale of the CTRS-R:L was significantly correlated only with the Negative Self-Esteem subscale of the CDI (r = .65), and the ADHD Index was significantly correlated only with the CDI Ineffectiveness subscale (r = .43). None of the CTRS-R:L subscales were significantly correlated with the CDI Interpersonal Problems subscale, and only one (Emotional Lability) was significantly correlated (r = .40) with CDI Negative Mood.

Conners (1997) also reports two studies in which children's and adolescents' scores on a task designed to assess vigilance or attention—the Continuous Performance Test (CPT; Conners, 1992, 1994) were correlated with parents' or teachers' ratings on the CRS-R. The CPT requires children to sit at a computer and to respond to stimuli as they appear on screen. It is repetitive and monotonous, and thus taxes children's attentional abilities. High scores on the CPT Overall Index are indicative of attention problems. The sample sizes for both of these studies were approximately 50. The mean age of children and adolescents in the study including parent ratings was 9.40 (SD = 1.98), while the mean age in the study including teacher ratings was 8.96 (SD = 1.68). Results from both studies indicated some expected significant correlations between these two types of measures, but other expected correlations were not significant (see Connors, 1997, p.134).

- The CPT Overall Index was significantly correlated with the CPRS-R:L DSM-IV Inattentive scale (r = .33, p < .05), and with the factor analytically derived scales Cognitive Problems/Inattention (r = .44, p < .05) and Psychosomatic (r = .37, p < .05), but expected significant correlations with other scales tapping hyperactivity and attention problems (e.g. Hyperactivity, CGI Restless-Impulsive, DSM-IV Hyperactive-Impulsive) were nonsignificant.
- The CPT Overall Index also had a significant correlation with the CTRS-R:L Cognitive Problems/Inattention scale (r = .35, p < .05), and was negatively correlated with teacher-rated Perfectionism (r = -.35, p < .05). Other expected significant correlations with other hyperactivity and attention problems subscales were nonsignificant (e.g. Hyperactivity, CGI Restless-Impulsive, DSM-IV Inattentive, DSM-IV Hyperactive-Impulsive).

Discriminant Validity

Conners (1997) presents evidence for the discriminant validity of the DSM-IV Symptoms scales in two ways (p.136).

- First, he determined the percentages of children and adolescents from the norming sample who met the DSM-IV criteria for diagnosis of ADHD, inattentive subtype, hyperactive-impulsive subtype, or combined subtype, based on CPRS-R:L and CTRS-R:L ratings. A child is considered to meet the criteria for diagnosis of ADHD inattentive or hyperactive-impulsive subtype if the parent or adult respondent reports that at least six of nine symptoms associated with the subtype are "very much true" of the child. If the child meets criteria for both subtypes (i.e., if the parent or teacher reports that at least six of the nine symptoms associated with each subtype are "very much true" for the child), the child receives the classification of ADHD, combined subtype. The percentages of the norming sample who met the diagnostic criteria for one of the three ADHD subtypes were 3.85 percent based on teacher ratings, and 2.3 percent based on parent ratings. Conners reports that these percentages are consistent with the expected percentages in the population of school-age children (3 to 5 percent), as reported in the DSM-IV.
- Second, he compared mean subscale scores of the children identified as meeting diagnostic criteria for ADHD (as described above) with mean subscale scores for randomly selected non-ADHD children, matched for sex and age.
    - CPRS-R:L analyses included 57 ADHD children (42 male, 15 female) with a mean age of 9 years, 6 months (*SD* = 3 years, 4 months) and a matched sample of 57 non-ADHD children. Results of *t*-tests indicated that the ADHD group had significantly higher scores than the non-ADHD children on all subscales except Perfectionism.
    - CTRS-R:L analyses included 76 ADHD children (56 male, 17 female) with a mean age of 8 years, 9 months. (*SD* = 2 years, 9 months) and a matched sample of 76 non-ADHD children. Consistent with the CPRS-R:L results, the ADHD group had significantly higher scores than the non-ADHD children on all subscales except Perfectionism.

To further assess the discriminant validity of the CPRS-R:L and the CTRS-R:L, Conners compared subscale scores for three groups of children: 1) children and adolescents who had received an independent diagnosis of ADHD, 2) children who had been identified by a psychologist or psychiatrist as having "emotional problems," and 3) a nonclinical group randomly selected from the norming sample to match the independently-diagnosed ADHD group as closely as possible on age, sex, and ethnicity (see Connors, 1997, p.137-138).

- CPRS-R:L analyses included 91 children and adolescents (70 male, 21 female) with a mean age of 10 years, 3 months (SD = 3 years, 5 months) in the ADHD group, a matched nonclinical group of 91 children and adolescents, and an emotional problems group including 55 children and adolescents (42 male, 13 female), with a mean age of 11years, 8 months (*SD* = 2 years, 10 months). Age was included as a covariate in all analyses because of the older mean age of the emotional problems group. Results were consistent with expectations and were interpreted by Connors as being indicative of symptom specificity of the CPRS-R:L subscales.

  - The nonclinical group was significantly lower than the emotional problems group on all subscales, and lower than the ADHD group on all subscales except Perfectionism.
  - The ADHD group was significantly higher than the emotional problems group on subscales reflecting attentional problems and hyperactivity, including Cognitive Problems, Hyperactivity, Restless-Impulsive, CGI Total Score, DSM-IV Inattentive, DSM-IV Hyperactive-Impulsive, DSM-IV Total, and the ADHD Index.
  - The emotional problems group was significantly higher than the ADHD group on the Oppositional, Perfectionism, and Social Problems subscales.
- CTRS-R:L analyses included 154 children and adolescents (122 male, 32 female) with a mean age of 10 years, 5 months (SD = 3 years, 6 months) in the independently-diagnosed ADHD group, a matched nonclinical group of 154 children and adolescents, and an emotional problems group including 131 children and adolescents (105 male, 26 female), with a mean age of 12 years, 7 months (*SD* = 2 years, 11 months). Age was again included as a covariate in all analyses because of the older mean age of the emotional problems group. Results were consistent with expectations and with those reported for parent ratings.
  - The nonclinical group was significantly lower than the emotional problems group on all subscales and lower than the ADHD group on all subscales except Social Problems.
  - The independently-diagnosed ADHD group was significantly higher than the emotional problems group on subscales reflecting attentional problems and hyperactivity, including Cognitive Problems, Restless-Impulsive, CGI Total Score, DSM-IV Inattentive, DSM-IV Hyperactive-Impulsive, DSM-IV Total, and the ADHD Index. The two groups were not significantly different on the Hyperactivity subscale, however.
  - The emotional problems group was significantly higher than the ADHD group on the Oppositional, Perfectionism, Emotional Lability, and Social Problems subscales.

Reliability/Validity Information from Other Studies

- None found.

Comments

- With respect to internal consistency reliability, information provided by Conners (1997) indicated strong internal for most of the scales for both males and females, with the exception of two scales: CGI Emotional Lability for younger females on the CPRS-R:L, and Social Problems for younger females on the CTRS-R:L. Information from the CTRS-R:L suggests somewhat lower internal consistency for teachers' ratings of girls than boys within the youngest age group.
- Moderate to high correlations between ratings obtained between 1½ and 2 months apart provided support for the test-retest reliability of the CPRS-R:L and CTRS-R:L. Test-retest reliabilities of the CPRS-R:S and the CTRS-R:S were not directly examined, although estimates for short form scales derived from long form versions of the measures indicated high test-retest reliabilities.

- With respect to inter-rater agreement, correlations between parallel parent- and teacher-report subscales varied widely for both males and females, indicating low levels of agreement between raters for both males and females for some child behaviors, including Perfectionism, Emotional Lability, and Oppositional behavior, and more moderate or high levels of agreement for other types of behavior, including Cognitive Problems/Inattention, Restless-Impulsive behavior, and the ADHD Index (see Connors, 1997, p.128-129). These results may be expected give the different contexts in which parents and teachers observe children and may speak to the value of having multiple perspectives on children's behavior, particularly when attempting to assess children for the presence of absence of behavioral and emotional problems.
- With respect to convergent (factorial) validity, correlations among subscales of the CPRS-R and the CTRS-R long and short forms support Conners' conclusion that the factor analytically derived subscales tap distinctive behavioral and emotional problem areas, although strong correlations (i.e., above .50) between some subscales also suggests that behavioral and emotional problems as assessed by the subscales of the various forms of the CRS are not independent.
- The evidence of convergent validity presented by Conners (1997) is fairly weak. The CDI and the Conners' Teacher and Parent Rating Scales do not have scales that are truly parallel. Expected associations between the CRS-R measures and the CPT were not consistently significant. Further, the small sample sizes of the reported studies resulted in limited statistical power, and some correlations that might have been both significant and meaningful in larger samples were consequently nonsignificant.
- With respect to discriminant validity, evidence presented by Conners (1997) of differences between independently-identified groups provides some support for the validity of the CRS-R subscales. Evidence of mean differences on the CRS-R subscales between children who did or did not meet diagnostic criteria for a diagnosis of ADHD based on the CRS-R was less compelling, given that the diagnosis and the subscale scores were derived from the same measure.
- Perfectionism was repeatedly found to have the lowest associations with other subscales from both the CPRS-R:L and the CTRS-R:L. Based on information provided by Conners (1997), it is clear that Perfectionism as tapped by these measures has little association with other emotional and behavioral problems tapped by the CRS-R subscales, although the reasons for these low associations are unclear.
- Although full age ranges of the validity studies were not presented, it appears that the validity studies conducted by Conners included few, if any, preschool children or young school-age children. Additional studies need to be conducted to determine the validity of these measures for younger children.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- The CRS-R scales are fairly recent revisions of earlier measures, and few reports have yet appeared using these scales in their current form. Earlier versions of CRS-R scales have been used in research to identify children with and without ADHD or hyperactivity

problems (e.g., Cohen, 1983). Other studies have used Conners' scales in pretest post-test designs to determine whether drug, behavior modification, or other therapies differentially improve the behavior of children independently diagnosed with ADHD (e.g., Pelham, Swanson, Furman, & Schwindt, 1996). Pelham, Swanson, Furman, & Schwindt (1996) found that the drug pemoline had an effect on academic performance as measured by an Abbreviated Conners' Teacher Rating Scale. This effect was measured two hours after taking the drug and was measured through seven hours after. However, no studies were found in which the CRS-R scales were used with general samples of children to detect changes in behavior resulting from interventions.

V. Adaptations of Measure

None found.

Early Childhood Measures: Social-Emotional

Early Childhood Measures: Social-Emotional

Devereux Early Childhood Assessment (DECA)

I. Background Information

Author/Source

Authors: P.A. LeBuffe & J.A. Naglieri

Source: LeBuffe, P.A., & Naglieri, J.A. (1999). *Devereux Early Childhood Assessment Program: Technical Manual*. Lewisville, NC: Kaplan Press.

Publisher: Kaplan Press
1310 Lewisville-Clemmons Rd.
Lewisville, NC 27023
800-334-2014
Website: www.kaplanco.com

Purpose of Measure

As described by instrument publisher:
The DECA is a nationally normed instrument designed to evaluate preschool children's social-emotional strengths that have been found in the developmental literature to be associated with resiliency. The authors suggest that the DECA can be used as an assessment to determine the needs of individual children, or to develop classroom profiles that may facilitate optimal classroom and instructional design.

Population Measure Developed With

- The DECA was developed over a two-year period between 1996 and 1998.
- The standardization sample for the Protective Factors component of the DECA was a nationally representative sample of 2,000 children (51 percent boys and 49 percent girls) aged 2 years through 5 years and 11 months of age, collected from all regions of the United States. Approximately half (983) of the children were rated by a parent or other family caregiver, and half (1,017) were rated by a teacher or childcare provider.
- Information on race and Hispanic origin are presented separately. In the Protective Factors sample, excluding children whose race was identified as "other," 76.3 percent of the children were white, 18.8 percent were black, 3.8 percent were Asian/Pacific Islander, and 1.0 percent were American Indian/Alaskan Native. These percentages closely approximate the distribution in the U.S. population. In the DECA sample, 10.7 percent of children were of Hispanic origin, close to the percentage reported in the U.S. population in 1995.
- A separate standardization sample was collected for the Behavioral Concerns component of the DECA. This sample included 1,108 children aged 2 years to 5 years 11 months (51 percent boys and 49 percent girls). Like the Protective Factors sample, this sample was collected from all regions of the United States. Half of the children (541) were rated by their parents, and half (567) were rated by preschool teachers.

- In the Behavioral Concerns sample, excluding children whose race was identified as "other," 79.9 percent of the children were white, 17.0 percent were black, 2.1 percent were Asian/Pacific Islander, and 1.0 percent were American Indian/Alaskan Native. As with the Protective Factors sample, these percentages closely approximate the distribution in the U.S. population. In this sample, 9.2 percent of children were of Hispanic origin, close to percentage reported in the U.S. population in 1995.

Age Range Intended For

Ages 2 years through 5 years.

Key Constructs of Measure

- *Protective Factors Scale*:
  - *Initiative:* Items tap the child's ability to think and act independently. Included are items involving preference for challenge and persistence.
  - *Self-Control:* Items reflect the ability to experience a range of emotions and to express emotions in appropriate ways.
  - *Attachment:* Items tap strong positive social bonds between the child and adult(s).
- *Behavioral Concerns*:
  - The 10 items on this scale reflect a number of problematic behaviors that may be exhibited by young children, including angry, aggressive, and destructive behavior and attention problems.

Norming of Measure (Criterion or Norm Referenced)

Norm referenced.

Comments

- Associated with the DECA instrument are curricular materials, including classroom strategies focusing on individual children and the classroom as a whole, as well as materials related to working with families. The foundation of this work is in the developmental literature on risk and resiliency (e.g. Garmezy, 1985; Werner & Smith, 1982), and the goal of the DECA program is to strengthen characteristics of children and their environments that promote resilience.
- The 37-item DECA focuses on positive behavioral dimensions that are believed to be important for successful functioning in school and other settings. A particular strength may be its inclusion of a scale that taps behaviors frequently included within the approaches to learning construct—the Initiative scale.
- The DECA would likely be insufficient to use alone, however, when detection of behavioral and emotional problems is needed or desired, as it contains only a single Behavioral Concerns scale that may differentiate children who are experiencing problems from those who are not, but does not provide a comprehensive picture of the particular types of difficulties that individual children—or groups of children—are experiencing.
- Concerns have been raised about the labels of scales, although not necessarily with the content. As a measure of child-based characteristics, use of the term "Protective Factors" does not necessarily match with its use in developmental literature, where the focus is

most frequently on positive characteristics of the child's family as well as the strength of support systems outside of the family. The Attachment subscale cannot adequately capture security of attachment as typically defined in the literature. Attachment, conceptualized as a dyadic construct, is most frequently assessed through observations of interactions between a child and a parent or caregiver. The DECA scale should perhaps rather be described as a measure of social responsiveness and sociability.

II. Administration of Measure

Who is the Respondent to the Measure?
Parents and teachers or childcare providers.

If Child is Respondent, What is Child Asked to Do?
Not applicable.

Who Administers Measure/Training Required?
Test Administration:

- The DECA is a brief (37-item) questionnaire that does not require any special training to administer. Training programs are available for the DECA program, including classroom practices and assessments. Users should be trained in the interpretation and use of standardized assessment instruments.

Data Interpretation:

- (Same as above.)

Setting (e.g., one-on-one, group, etc.)
Parents and other adults typically complete the DECA independently.

Time Needed and Cost
Time:

- Completion of the DECA takes approximately 10 minutes (per child).

Cost:

- Complete kit (includes assessment materials and materials for implementing the DECA program in classrooms): $199.95
- Technical Manual: $19.95
- Record forms: $39.95 for a pack of 40

Comments

- The DECA is among the easiest social-emotional assessments to administer and scoring is straightforward.

III. Functioning of Measure

Reliability Information from Manual

Internal Consistency

In the original Protective Factors standardization sample, internal consistency reliability of the Total Protective Factors scale was .91 for parent report, .94 for teacher report. Reliabilities of the Protective Factors subscales ranged from .76 for parent report Attachment to .90 for teacher report Initiative and Self-Control. Internal reliability of the Behavioral Concerns scale within the Behavioral Concerns standardization sample was .71 for parent report, .80 for teacher report (see LeBuffe & Naglieri, 1999, p. 16).

Test-Retest Reliability

Test-retest reliabilities (correlations) were obtained with a sample of 26 children (42.3 percent boys, 57.7 percent girls) who were rated twice by the same parent, and a separate sample of 82 children (48.8 percent boys, 51.2 percent girls) who were rated twice by the same teacher. The time interval between ratings ranged from 1 to 3 days. Across this very short time interval, correlations for parent ratings ranged from .55 for both Attachment and Behavioral Concerns, to .80 for Initiative. Test-retest correlations of teacher ratings ranged from .68 for Behavioral Concerns to .94 for the Total Protective Factors scale. As with internal consistency, test-retest reliabilities were consistently higher for teacher reports than for parent reports (see LeBuffe & Naglieri, 1999, p. 18).

Interrater Reliability

Independent ratings by up to four raters (two parents and two teachers) were collected on a sample of preschool children. All ratings were conducted on the same day. A total of 62 children (48.4 percent boys, 51.6 percent girls) were rated by two parents, 80 children (47.5 percent boys, 52.5 percent girls) were rated by two teachers or teacher's aides, and 98 children (52.0 percent boys, 48.0 percent girls) were rated by at least one parent and one teacher or teacher's aide (see LeBuffe & Naglieri, 1999, p. 20).

- Interrater reliability for pairs of teachers and teacher's aides ranged from .57 for Attachment to .77 for Self-Control.
- Correlations between mothers' and fathers' ratings were considerably lower, ranging from .21 (not significant) for Total Protective Factors to .44 for Behavioral Concerns.
- Correlations between parent and teacher ratings were also lower than those between ratings by teachers and aides, ranging from .19 (not significant) for Attachment to .34 for Initiative.

Validity Information from Manual

Criterion Validity

A sample of 95 children (66 percent boys, 34 percent girls) with identified emotional or behavioral problems (i.e., who had received a psychiatric diagnosis, who were receiving mental health services, or who were asked to leave a child care program due to their behavior) were compared with a matched community sample of 86 children (67 percent boys, 33 percent girls) with no identified emotional or behavioral problems.

- As was predicted, mean standardized (*T*) scores on the Protective Factors scales were consistently higher in the community sample (*T* scores ranging from 47.0 for Total Protective Factors to 49.1 for Self-Control) compared to the identified sample (*T* scores ranging from 38.5 for Total Protective Factors to 41.9 for Attachment), while the identified sample children on average had higher scores than did the community sample children on the Behavioral Concerns scale (*T* scores of 65.4 and 55.7, respectively). All mean differences were significant (see LeBuffe & Naglieri, 1999, p. 26).
- As a further test of criterion validity, LeBuffe and Naglieri (1999) predicted that children with standardized scores within the range considered to be of concern (i.e., *T* scores less than 40 on the Total Protective Factors Scale or above 60 on the Behavioral Concerns Scale) would be more likely to be in the group of children with identified problems. Results of these analyses supported the validity of the DECA in identifying children with potential problems: A total of 67 percent of children in the identified group had Total Protective Factors *T* scores of 40 or below; 29 percent of children in the community sample had scores that low. For the Behavioral Concerns scale, 78 percent of children in the identified sample had *T* scores of 60 or higher, while 35 percent of children within the community sample had scores that high (see LeBuffe & Naglieri, 1999, p. 28).

Reliability/Validity Information from Other Studies

- None found.

Comments

- Internal consistency reliability was somewhat higher for teacher ratings than for parent ratings, although all reported alphas were high.
- Test-retest correlations across a 1- to 3-day interval were also high for all DECA scales. Correlations were consistently higher for teacher ratings than for parent ratings, however, and correlations as low as .55 (for parent report of children's Attachment and Behavioral Concerns) across such a brief interval may indicate that parents are responding somewhat differently at the two assessments. The reasons for this are unclear but may include relatively systematic testing effects (e.g., greater familiarity with the DECA following the first administration, increased focus on and evaluation of children's behaviors following the first administration).
- As might be expected, information on interrater reliability presented by LeBuffe and Naglieri (1999) indicates that DECA ratings are more consistent when children are observed by two raters at the same time and in the same setting (i.e., the preschool classroom) than when raters observe children in different contexts (i.e., home and preschool).
- Data presented by LeBuffe and Naglieri (1999) provides support for the criterion-related validity of DECA Behavioral Concerns and Protective Factors scales.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- The 19 items comprising the Self-Control and Initiative scales of the DECA were included in an evaluation of the effects of viewing *Dragon Tales*, an educational television program produced with funding from the Corporation for Public Broadcasting through the U.S. Department of Education (Rust, 2001). The program is targeted at children between 2 and 6 years of age and is designed to help children learn positive strategies for dealing with social, emotional, physical, and cognitive challenges in their lives. The evaluation included three studies, two conducted in school, the third involving in-home viewing. In one of the school-based studies, a pretest/post-test design was used with a sample of 340 4- and 5-year-olds to compare a group of children who watched Dragon Tales in school daily for several weeks with a group of children who watched another program, *Between the Lions*, that is primarily designed to promote literacy. DECA evaluations were completed by teachers, parents, and researchers. Item-level analyses of averaged teacher, parent, and researcher ratings on the DECA indicated that the group of children who watched Dragon Tales demonstrated significantly increased scores from pretest to post-test on six DECA items tapping sharing, cooperating, leadership in play interactions with peers, and choosing challenging tasks, relative to changes in the Between the Lions group. Fewer significant differences were found for teacher or researcher ratings alone, and no significant differences were found between the two groups of children based on parent reports.
- The DECA Program has been implemented in Head Start programs, and the DECA assessment instrument has been or is being used (with or without the associated Program) in evaluations of preschool program effectiveness across the country. The Devereux Foundation web site (www.devereuxearlychildhood.org) has several brief reports regarding use of the instrument to assess program effectiveness.

V. Adaptations of Measure

Spanish Version of the DECA

Description of Adaptation

The Spanish-language version of the DECA was developed through a process of having a professional translator with experience in child development create a version, which was then evaluated by three bilingual English-Spanish speakers who back-translated the Spanish version into English. Minor changes in the Spanish version were made on the basis of these back-translations.

Psychometrics of Adaptation

Equivalence of the English and Spanish versions of the DECA was assessed by asking 92 bilingual individuals (44 parents and 48 teachers; 49 percent Mexican, 24 percent Puerto Rican, 27 percent other) to each rate a child with both the English and the Spanish versions, with the order in which the ratings were done counterbalanced across raters.

Paired sample *t*-tests indicated no significant differences between ratings on the DECA scales and subscales across the English and Spanish versions for either parents or teachers.

Study Using Adaptation

None found.

Early Childhood Measures: Social-Emotional

Early Childhood Measures: Social-Emotional

Social Competence and Behavior Evaluation (SCBE) – Preschool Edition

I. Background Information

Author/Source

Source: LaFreniere, P. J., & Dumas, J. E. (1995). *Social Competence and Behavior Evaluation—Preschool Edition (SCBE)*. Los Angeles, CA: Western Psychological Services.

Publisher: Western Psychological Services (WPS)
12031 Wilshire Blvd.
Los Angeles, CA 90025-1251
Phone: 800-648-8857
Website: www.wpspublish.com

Purpose of Measure

As described by instrument publisher:
To assess emotional adjustment and social competence in children aged 2 years, 6 months through 6 years, 6 months. The SCBE is designed to assess positive aspects of social adjustment, to differentiate among different types of emotional and behavioral adjustment difficulties, and to be sensitive to changes across time and treatment. As described in the test manual, "The primary objective of the SCBE is to describe behavioral tendencies for the purposes of socialization and education, rather than to classify children within diagnostic categories" (LaFreniere & Dumas, 1995, p. 1).

Population Measure Developed With

- The SCBE was previously entitled the Preschool Socio-Affective Profile (PSP). Early work with the SCBE was conducted with French-Canadian samples (French-speaking) in Montréal, Canada. Following an initial pilot study, the first reliability study was conducted with a sample of 979 preschool children (458 girls, 521 boys) enrolled in 90 urban preschool classrooms.
- The SCBE was subsequently translated into English and standardization research was conducted on a large sample of children attending 100 different preschool classrooms located in four Indiana cities (Indianapolis, Lafayette, Frankfort, and Logansport) and two Colorado cities (Denver and Boulder). A total of 1,263 children were included in the sample (631 girls, 632 boys). Children ranged in age from 2 years, 6 months through 6 years, 6 months, with the largest percentage of children being between 4 years, 7 months and 5 years, 6 months of age (41.7 percent of the sample). Parent education levels were relatively low in this sample, with 26.7 percent having less than 12 years of education, 43.0 percent having a high school diploma, 16.6 percent having some college training, and 13.6 percent having four years of college or more. Compared with U.S. Census figures for adults ages 25-44 (U.S. Bureau of the Census, 1991, as cited in LaFreniere & Dumas, 1995), the percentage of parents having less than 12 years of

education is approximately twice the national percentage (13.0 percent), and the percentage of parents who had four or more years of college was approximately half the national percentage (26.0 percent). The majority of children were white (68.4 percent), 20.6 percent were black, 7.3 percent were Hispanic, and 3.6 percent were Asian/Pacific Islander. For comparison, the national percentages of children ages 9 and below were 80.0 percent White, 15.0 percent black, 11.0 percent Hispanic, and 2.0 percent Asian (U.S. Bureau of the Census, 1991, as cited in LaFreniere & Dumas, 1995).

Age Range Intended For

Ages 2 years, 6 months through 6 years, 6 months.

Key Constructs of Measure

Basic Scales

There are a total of eight different Basic Scales, each with a "negative pole" tapped by five items, and a "positive pole" tapped by five items.

- *Overall adjustment scales*:
  - *Depressive-Joyful:* Taps the child's characteristic mood.
  - *Anxious-Secure:* Items related to the child's sense of security in the classroom. Includes items relevant to motivation, self-confidence, and approaches to learning.
  - *Angry-Tolerant:* The child's ability to effectively manage challenges and frustrating experiences typical of a preschool classroom.
- *Peer social interactions scales:*
  - *Isolated-Integrated:* The extent to which the child is part of the peer group, versus being socially isolated.
  - *Aggressive-Calm:* Taps the extent to which the child engages with peers in aggressive or prosocial ways, particularly in conflict situations.
  - *Egotistical-Prosocial:* Perspective-taking in interactions with peers.
- *Adult social interactions scales:*
  - *Oppositional-Cooperational:* The tendency to be appropriately compliant versus noncompliant in interactions with adults.
  - *Dependent-Autonomous:* The ability of the child to engage in independent activities, versus over-reliance on assistance and comfort from adult caregivers.

In addition, there are four *Summary Scales:*

- *Social Competence:* Summarizes the eight positive poles (Joyful, Secure, Tolerant, Integrated, Calm, Prosocial, Cooperational, and Autonomous; 40 items).
- *Internalizing Problems*: Summarizes the Depressive, Anxious, Isolated, and Dependent negative poles (20 items).
- *Externalizing Problems*: Summarizes the Angry, Aggressive, Egotistical, and Oppositional negative poles (20 items).
- *General Adaptation*: Summarizes both positive and negative poles (i.e., all 80 questionnaire items).

Norming of Measure (Criterion or Norm Referenced)

Norm referenced. No separate norms were developed for different ages. The authors suggest that scores corresponding to the 10th percentile and the 90th percentile of the norming sample are

clinically significant thresholds identifying children who demonstrate unusually poor or positive adjustment.

Comments

- This measure appears to be among the most oriented toward examining individual differences in children generally, rather than beginning as an instrument for identifying children with specific emotional and behavioral disorders. The balance between positive and negative characteristics is exactly even, a unique characteristic of this measure.
- The sample was not specifically designed to be representative of the entire U.S. population of preschool-age children. The sample was selected from a few cites that did not represent all regions of country. Parent education level was relatively low, and there was an overrepresentation of black and Asian children relative to percentages in the national population.

II. Administration of Measure

Who is the Respondent to the Measure?
Teacher.

If Child is Respondent, What is Child Asked to Do?
N/A.

Who Administers Measure/ Training Required?
Test Administration:

- The SCBE is a teacher-report instrument, and researchers or clinicians do not require special training to use the SCBE. The authors indicate that teachers who complete the SCBE should be well-acquainted with the child. Thus, the SCBE should not be completed by a teacher immediately upon the child's entry into the preschool classroom. The authors further suggest that response accuracy may be improved when teachers are allowed to familiarize themselves with the SCBE items several weeks prior to a formal SCBE evaluation session.

Data Interpretation:

- Although no special training is required to administer the SCBE, interpretation of individual children's profiles for assessment purposes requires clinical training.

Setting (e.g., one-on-one, group, etc.)
Teachers complete the SCBE independently.

Time Needed and Cost
Time:

- The SCBE takes approximately 15 minutes per student to complete.

Cost:

- Starter kit (including Manual and 25 scoring sheets): $79.95
- Manual: $45.00

Comments

- There is currently no published parent report version of the SCBE.

Functioning of Measure

Reliability Information from Manual

Internal Consistency

In the French-Canadian sample, Cronbach's alphas for the eight basic scales ranged from .79 to .91. In the American standardization sample, alphas for the eight basic scales ranged from .80 to .89 in the Indiana subsample, and from .80 to .88 in the Colorado subsample (see LaFreniere, & Dumas, 1995, p. 42).

Test-Retest Reliability

In the French-Canadian sample, 29 students were reassessed after a two-week interval, and again six months later. Pearson correlations for the two-week interval ranged from .74 to .87. After six months, test-retest correlations ranged from .59 to .70. Test-retest reliability was not evaluated in the American standardization sample (see LaFreniere, & Dumas, 1995, p. 34).

Interrater Agreement

Interrater reliability was calculated using different teachers' independent evaluations of children. Results from the original French-Canadian sample indicated Spearman-Brown correlations for the eight basic scales ranging from .72 to .89.

In the American standardization sample, interrater agreement was assessed for the Indiana subsample only (824 children). Interrater reliability was calculated for two preschool teachers independently evaluating the same child at the same time. Consistent with findings from the French-Canadian sample, Spearman-Brown correlations ranged from .72 to .89 for the eight basic scales (see LaFreniere, & Dumas, 1995, p. 42).

Validity Information from Manual

Construct Validity

Construct validity was assessed similarly in the French-Canadian sample and in the American standardization samples. First, scores were constructed for each positive and negative pole of each of the basic scales (for a total of eight positive poles, or "item clusters" and eight negative poles, or "item clusters"). Principle components analyses were then conducted with these positive and negative item clusters. According to LaFreniere and Dumas (1995), results were consistent across all three samples (French-Canadian, Indiana, and Colorado), supporting the hypothesized three major constructs tapped by the SCBE: Social Competence (including factor loadings ranging from .58 to .81 for all eight positive pole item clusters), Externalizing Problems (including positive loadings ranging from .83 to .89 for Angry, Aggressive, Egotistical, and Oppositional item clusters and weaker negative cross-loadings ranging from -.39 to -.55 for corresponding positive pole item clusters), and Internalizing Problems (including positive loadings ranging from .75 to .84 for Depressive, Anxious, Isolated, and Dependent item clusters

and weaker negative cross-loadings ranging from -.53 to -.61 for corresponding positive pole item clusters; see LaFreniere, & Dumas, 1995, p. 34).

Convergent and Discriminant Validity

Convergent and discriminant validity were assessed in the French-Canadian sample by examining correlations between SCBE negative item clusters (Anxious, Isolated, Aggressive) and scales (Internalizing and Externalizing Problems) and corresponding scales from the Achenbach Child Behavior Checklist – Teacher Report Form (CBCL-TRF; Anxiety, Withdrawal, Aggression, Internalizing, and Externalizing; Achenbach, 1997), with the expectation that scales/item clusters designed to tap the same constructs would correlate more highly than would scales/item clusters tapping different constructs (e.g., SCBE Internalizing would correlate more highly with CBCL-TRF Internalizing than with CBCL-TRF Externalizing). Positive adaptation is not represented in the CBCL-TRF and thus no convergent or discriminant validity analyses could be conducted with the CBCL-TRF for the positive scales and item clusters from the SCBE. In general, the findings reported by LaFreiniere and Dumas demonstrate an expected pattern of associations, with SCBE and the CBCL-TRF scales tapping the same types of behavioral and emotional problems (i.e., internalizing or externalizing problems) being more highly associated with each other than with scales tapping different types of problems (see LaFreniere, & Dumas, 1995, p. 43).

- For boys, The SCBE Anxious Scale was correlated .48 with the CBCL-TRF Anxiety scale, .48 with the CBCL-TRF Withdrawal scale, and .52 with the CBCL-TRF Internalizing scale, while correlations with Aggression and Externalizing scales from the CBCL-TRF were low (.01 and .15, respectively). For girls, a similar pattern emerged. The SCBE Anxious Scale was correlated .40 with the CBCL-TRF Anxiety scale, .37 with the CBCL-TRF Withdrawal scale, and .43 with the CBCL-TRF Internalizing scale. Correlations with Aggression and Externalizing scales from the CBCL-TRF were .10 and .19, respectively.
- For boys, The SCBE Isolated Scale was correlated .58 with the CBCL-TRF Withdrawal scale, .51 with the CBCL-TRF Anxiety scale, and .59 with the CBCL-TRF Internalizing scale, while correlations with Aggression and Externalizing scales from the CBCL-TRF were negative and low (-.11 and -.01, respectively). For girls, the SCBE Isolated Scale was correlated .53 with the CBCL-TRF Withdrawal scale, .30 with the CBCL-TRF Anxiety scale, and .47 with the CBCL-TRF Internalizing scale. Correlations with Aggression and Externalizing scales from the CBCL-TRF were low (-.01 and .09, respectively).
- For both boys and girls, the SCBE Aggressive scale had significant correlations with the CBCL-TRF Aggression scale (.53 and .63 for boys and girls, respectively), and with CBCL-TRF Externalizing (.49 and .61 for boys and girls, respectively), while correlations with Anxiety, Withdrawal, and Internalizing scales were negative, low and nonsignificant (ranging from -.01 to -.12).
- For boys, The SCBE Internalizing Problems Scale was correlated .63 with the CBCL-TRF Internalizing scale, .57 with the CBCL-TRF Anxiety scale, and .60 with the CBCL-TRF Withdrawal scale; correlations with Aggression and Externalizing scales from the CBCL-TRF were lower (.13, and .27, respectively). For girls, the SCBE Internalizing Problems Scale was correlated .53 with the CBCL-TRF Internalizing scale, .50 with the CBCL-TRF Anxiety scale, and .45 with the CBCL-TRF Withdrawal scale. Correlations

with Aggression and Externalizing scales from the CBCL-TRF were lower (.20 and .29, respectively).

- For boys, The SCBE Externalizing Problems Scale was correlated .64 with the CBCL-TRF Externalizing scale, and .68 with the CBCL-TRF Aggression scale. Correlations with Anxiety, Withdrawal, and Internalizing scales from the CBCL-TRF were low and nonsignificant (.00, -.07 and -.03, respectively). For girls, the SCBE Externalizing Problems Scale was correlated .66 with the CBCL-TRF Externalizing scale, and .71 with the CBCL-TRF Aggression scale. Correlations with Anxiety, Withdrawal, and Internalizing scales from the CBCL-TRF were lower (-.03, -.20 and .12, respectively).

In addition to examining correlations between parallel scales on the SCBE and the CBCL-TRF, the authors also point to a relatively small association (a correlation of .28) between Internalizing and Externalizing problems scales on the SCBE as an indication of the discriminant validity of these two scales, and contrast this with a higher correlation (.60) found between CBCL-TRF Internalizing and Externalizing scales. According to Dumas and LaFreniere (1995), "This significantly greater orthogonality of the SCBE assessments of externalizing and internalizing problems is thought to be optimal for current research in developmental psychopathology investigating specific etiologies and sequelae of early patterns of disorder" (p. 43).

Criterion Validity

LaFreniere, Dumas, Capuano, and Dubeau (1992) conducted a study examining associations between PSP (SCBE) scores, and classifications based on those scores, and outcomes expected to be associated with social competence and behavioral problems, including assessments of interaction with peers, peer acceptance, and peer rejection. These researchers found that children classified as anxious-withdrawn based on SCBE scores were more likely than were other children to be socially isolated, although not necessarily rejected by their peers. Children classified as angry-aggressive, in contrast, were the most interactive with their peers, but they were also the most likely to be peer-rejected. Children in the highly socially competent group were the most well-liked by their peers, while children in the average group were intermediate between the angry-aggressive and socially competent children in terms of their peer acceptance.

Reliability/Validity Information from Other Studies

- None found.

Comments

- Information provided by Dumas and Lafreniere (1995) is generally supportive of the reliability and validity of the SCBE. The relatively low correlation between Internalizing and Externalizing Problems scales may be a particular strength of this measure.

Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- LaFreniere and Dumas (1992) found predicted associations between children's classification as Competent, Average, and Anxious based on their SCBE scale scores and both child and maternal behavior during a problem-solving task. Competent children

exhibited more positive affect and more compliance than did Anxious children. More significant differences were found between the groups on maternal affect and behavior than on child affect and behavior, however, possibly indicating effects of the social environment on the development of children's behavior problems and social competence. Mothers of Competent children displayed more positive affect and behavior, and less negative affect, aversive behavior, and commands than did mothers of Anxious children. Mothers of Average children tended to fall between the other two groups on these characteristics. Mothers of Anxious children were likely to exhibit negative, aversive responses to their children's affect and behaviors even when children were compliant. Evidence that these reactions were transactional in nature, and not purely a maternal characteristic, was presented in an additional study (Dumas & LaFreniere, 1993) in which mothers of Anxious and Aggressive children were found to exhibit aversive responses with their own children, but not with unfamiliar children.

- **Intervention Study:** Capuano and LaFreniere (1993; LaFreniere & Capuano, 1997) used a pretest-posttest design to examine changes in children's social competence and behavior problems following a six-month intervention (20 sessions) that included parent education regarding children's developmental needs, child-oriented interaction during mother-child play, and problem behavior modification and parenting skills, as well as working with mothers to build a social support network. Changes in children's SCBE scores from pretest to posttest indicated that the treatment resulted in significant reductions in internalizing symptoms and significant increases in social competence.

Comments

- The SCBE is a relatively new measure that has been available in standardized format only since 1995. However, results both pre- and post-standardization are promising, particularly given the consistency in results across French-Canadian and U.S. samples.

Adaptations of Measure

SCBE-30

Description of Adaptation:

LaFreniere and Dumas (1996) have developed an abbreviated version of this teacher-report measure including only 30 items. The SCBE-30 includes three scales, each with 10 items, that parallel the three basic scales from the full 80-item measure (the SCBE-80): Social Competence, Anger-Aggressive (parallel to Externalizing Problem Behaviors), and Anxiety-Withdrawal (parallel to Internalizing Problem Behaviors).

In addition, a parent-report version of the SCBE-30 has been constructed with only very minor wording differences from the teacher-report version.

Psychometrics of Adaptation:

LaFreniere and Dumas (1996) present a description of the construction of the SCBE-30. SCBE-80 assessments were collected for four samples of preschool children: 910 children from 80 preschool classrooms in Montréal, Canada, 854 children from 50 classrooms in Indianapolis and Lafayette, Indiana, 439 children from 30 classrooms in Denver and Boulder, Colorado, and 443

children from 20 classrooms in Bangor and Orono, Maine. The 80 items were initially reduced to 45, 15 from each of the three basic scales (Social Competence, Internalizing Problem Behaviors, and Externalizing Problem Behaviors), based on considerations of levels of item endorsement, interrater reliability, and internal consistency.

Factor analyses were then conducted with the remaining 45 items. Results of these analyses again supported a 3-factor solution. The 45 items were then reduced further by selecting the 10 items with the highest loadings on each factor. This was followed by separate factor analyses of the remaining 30 items for the four different samples, and two additional analyses, each including half of the Montréal sample, in order to establish the stability of the factor-scale structure. Results from all analyses supported the hypothesized factor structure. The names of the two Problem Behaviors scales were changed to more closely reflect the nature of the items retained from the SCBE-80.

Interrater reliability was examined with the Montréal, Indiana, and Maine samples. In all cases, Spearman-Brown correlations ranged from .78 to .91.

Internal consistency of the three scales was assessed in all four samples. Cronbach's alpha coefficients ranged from .77 to .92.

Test-retest reliability was assessed two weeks after the initial assessment in a smaller subsample of the Montréal sample (29 children, rated by two teachers). Similarly, Indiana (409 children rated by 16 teachers within the same academic year) and Maine (45 children rated by two teachers during different academic years) samples were assessed after 6 months. Pearson correlations for the two-week intervals ranged from .78 to .86. Correlations for the six-month intervals were predictably lower, ranging from .75 to .79 for the Indiana sample and .61 to .69 for the Maine sample.

In addition to factor analytic results, evidence for discriminant validity of the scales was seen in correlations ranging from .02 to .29 between the Anger-Aggression scale and the Anxiety-Withdrawal scale in all four samples. The Social Competence scale demonstrated higher, negative correlations with both Anger-Aggression (ranging from -.37 to -.58) and Anxiety-Withdrawal (ranging from -.30 to -.43).

Construct validity was also addressed by examining correlations between SCBE-30 scales and their parallel SCBE-80 scales. Correlations ranged from .92 to .97 across all samples. Conduct Disorder and Anxiety-Withdrawal measures were also obtained from the Revised Behavior Problem Checklist (RBPC; Hogan, Quay, Vaughn, & Shapiro, 1989). SCBE-30 Anger-Aggression was correlated .87 with the RBPC Conduct Disorder scale, while the two Anxiety-Withdrawal measures were correlated .67.

Study Using Adaptation:
LaFreniere et al. (2002) conducted a cross-cultural analysis of the SCBE-30 with a total of 4,640 preschool children in eight countries: Austria, Brazil, Canada, China, Italy, Japan, Russia, and the United States. Results supported the structural equivalence of the SCBE-30 across all samples. Some cross-cultural differences in age trends in the prevalence of behavior problems

were evident, while a trend for increasing social competence across the preschool years was evident in all samples.

Spanish Version of the SCBE

Description of Adaptation:
Dumas, Martinez, and LaFreniere (1998) translated the SCBE for use with monolingual and bilingual Spanish-speaking preschool teachers.

Psychometrics of Adaptation:
Multiple bilingual Spanish speakers were recruited to participate in translating the SCBE into Spanish, and a different set of translators were recruited to back-translate the Spanish version to ensure accuracy. Translators were from multiple Spanish-speaking cultural backgrounds (Cuban, Puerto Rican, Mexican, Argentinean, Colombian, and Spanish), thus creating a measure that uses a sufficiently generic version of Spanish to be useful across a variety of cultures.

Test-retest reliability was assessed with a sample of 225 children with bilingual preschool teachers of Cuban background in Miami, Florida. Internal consistency was assessed with this same sample as well as with two additional samples of children with bilingual or monolingual teachers, one collected in Valencia, Spain (242 preschoolers), and the other collected in Houston, Texas (172 preschoolers). The factor structure of the newly translated SCBE was also assessed in the latter two samples. Results of all analyses in all three settings were consistent with each other and with results from French Canadian and U.S. English-speaking samples.

Study Using Adaptation:
None found.

Comments

- The development of a parent-report SCBE measure may be an important addition to this work. At this time, however, there have been no published reports involving the reliability or validity of this new measure.

Early Childhood Measures: Social-Emotional

Early Childhood Measures: Social-Emotional

Social Skills Rating System (SSRS)

III. Background Information

Author/Source

Source: Gresham, F.M., & Elliott, S.N. (1990). *Social Skills Rating System: Manual*. Circle Pines, MN: American Guidance Service.

Publisher: American Guidance Service (AGS)
4201 Woodland Rd.
Circle Pines, MN 55014-1796
Phone: 800-328-2560
Website: www.agsnet.com

Purpose of Measure

This summary focuses on the parent and teacher forms of the SSRS. There is a student form that can be administered to elementary and secondary students.

As described by instrument publisher:
The SSRS is designed to be an assessment of children's social behaviors, including strengths and weaknesses. The SSRS can be used for screening and identification of children suspected of having serious social behavior problems and to provide information that can assist in planning individualized interventions. Gresham and Elliott (1990) further indicate that the SSRS can be used to evaluate the social skills of groups, such as classrooms. Such group evaluation could be useful for designing classroom activities.

Population Measure Developed With

There was a national standardization sample for the elementary school version. Normative information for preschoolers was derived from a national "tryout" sample that received preliminary versions of the SSRS. Information from the tryout sample was used to develop the final, published version of the SSRS that was administered in the standardization study. No additional preschoolers were included in the standardization sample.

- *Characteristics of the preschool sample (from the national tryout study):*
  - There were 193 parent ratings of preschoolers, and 212 preschoolers were rated by 34 different teachers.
  - No information was provided by Gresham and Elliott (1990) on the demographic make-up of the tryout sample, other than indicating that children included in the sample were obtained from sites in 9 states.
- *Characteristics of elementary school age standardization sample:*
  - A total of 1,021 students were rated by 208 teachers, and 748 students were rated by a parent or guardian.
  - The sample was obtained from 20 sites located in 18 states across the country.

- The sample included children enrolled in special education classes as well as special needs students enrolled in regular classrooms. A total of 19.5 percent of the elementary school age children in the sample were classified as having some form of handicapping condition (predominantly learning disabilities).
- Minorities were somewhat underrepresented in the parent form sample, relative to the U.S. population. White parents made up 82.5 percent of the sample, 10.7 percent were black, 5.3 percent were Hispanic, and 1.5 percent were other minorities.
- Parents were somewhat more highly educated, on average, than the general U.S. population. Educational attainment was less than high school education for 8.8 percent of parents, 34.2 percent were high school graduates, 30.9 percent had some college or technical training, 26.1 percent had four or more years of college.
- No information on family income was presented for the sample.

Age Range Intended For

- Ages 3 years through 5 years (Preschool Forms).
- Grades K through 6 (Elementary Forms).

Key Constructs of Measure

Social Skills and Problem Behaviors are the major domains assessed by Parent and Teacher Report forms. Academic Competence is also included in the Teacher Report form. Each of the Social Skills items includes a second question regarding how important, in general, the adult feels that the behavior is. These importance ratings are seldom used in research.

- *Social Skills.* In addition to a Social Skills domain score, scores can be constructed for four separate subscales:
  - *Cooperation:* Includes helping, sharing, and compliance with rules.
  - *Assertion:* Taps initiation of behavior, questioning, self-confidence and friendliness.
  - *Self-control:* Focuses on appropriate responsiveness in conflict situations.
  - *Responsibility*: Included in the parent form only, this subscale taps ability to communicate with adults and respect for others and property.
- *Problem Behaviors*. In addition to a Problem Behaviors domain score, scores can be constructed for three separate subscales:
  - *Externalizing problems:* Includes verbal and physical aggression, noncompliance, and poor control of temper.
  - *Internalizing problems:* Taps anxiety, sadness and depression, loneliness, and low self-esteem.
  - *Hyperactivity:* Included in the Elementary Level forms only, this scale taps excessive movement and distractibility.

Norming of Measure (Criterion or Norm Referenced)

Norm referenced.

Comments

- A primary strength of the SSRS is its focus on positive social behaviors, and balance of subscales pertaining to positive and problem behaviors.

- Neither the tryout sample nor the standardization sample are truly nationally representative. Although little information is presented for the preschool tryout sample, lack of representativeness could be a cause for caution, particularly when attempting to interpret standardized scores based on norms for this age group.
- A number of concerns have been raised about the preschool version of the SSRS. It appears to be a downward extension of the elementary version, with some items of questionable appropriateness for preschoolers.

IV. Administration of Measure

Who is the Respondent to the Measure?

- Teacher or school personnel (teacher report form). Teacher report forms can be completed by teachers or other school personnel with sufficient experience interacting with the child.
- Parent or guardian (parent report form).

If Child is Respondent, What is Child Asked to Do?

Not applicable.

Who Administers Measure/Training Required?

Test Administration:

- Minimal training is required for administration of the questionnaires.
- Respondents (parents and teachers) need to be able to read at the third-grade level or above. The authors suggest that adults rating a child should have spent several days a week with the child for at least two months prior to rating.

Data Interpretation:

- Interpretation of responses should be done by professionals with training in psychological testing.

Setting (e.g., one-on-one, group, etc.)

One-on-one or independent. Typically, teacher and parent respondents complete questionnaires independently, although the questionnaire is sometimes administered to parents in an interview format (e.g. Pedersen, Worrell, & French, 2001).

Time Needed and Cost

Time:

- Administration of both the Parent and Teacher forms takes 15-25 minutes per child.

Cost:

- Scannable format questionnaire: $39.95 for packets of 25 forms
- Software for scoring and reporting, plus manual, ranges from $249.95 to $999.95 (depending in part upon whether group report capability is needed).

Comments

- The administration of SSRS forms is relatively straightforward. The ability to obtain parallel kinds of information about the child from multiple sources is a strength of the SSRS.

V. Functioning of Measure

Reliability Information from the Manual

The authors present reliability and validity information based on the tryout sample for the preschool forms, and on the standardization sample for the elementary form (Gresham & Elliott, 1990; Elliott, Barnard, & Gresham, 1989).

Internal consistency

For the teacher report, alpha coefficients for the social skills subscales (Cooperation, Assertion, Self-control, and Responsibility) ranged from .90 to .91 for preschool and .86 to .92 for elementary. Alphas for the Social Skills domain scale were .94 for both preschool and elementary forms. Alphas for Problem Behaviors subscales (Externalizing, Internalizing, and Hyperactivity) ranged from .74 to .85 for preschool, and from .78 to .88 for elementary. The Problem Behaviors domain scale alphas were .82 and .88 for preschool and elementary school forms, respectively (see Gresham, & Elliot, 1990, p. 109).

For the parent report, alpha coefficients were somewhat lower. Alphas for the social skills subscales ranged from .75 to .83 for preschool and .65 to .80 for elementary. For both age levels, Self-Control demonstrated the highest internal consistency, while Responsibility had the lowest consistency. Alphas for the Social Skills scale were .90 and .87 for preschool and elementary forms, respectively. Alphas for problem behavior subscales ranged from .57 (Internalizing) to .71 (Externalizing) for preschool, and from .71 to .77 for elementary. Problem Behaviors scale score alphas were .73 and .87 for preschool and elementary school forms, respectively (see Gresham, & Elliot, 1990, p. 109).

Test-retest reliability

Test-retest reliability information was not obtained for the preschool forms. For the elementary age version, with a four-week interval between ratings, correlations between teacher ratings for the social skills subscales ranged from .75 to .88, and the correlation for Social Skills scale scores was .85. For problem behaviors, correlations ranged from .76 to .83 for the subscales, and the across-time correlation for the Problem Behaviors scale was .84 (see Gresham, & Elliot, 1990, p. 111).

Across the same time interval, correlations between parent ratings for the elementary version of the social skills subscales ranged from .77 to .84, and the correlation for Social Skills scale scores was .87. For problem behaviors, correlations ranged from .48 to .72 for the subscales, and the across-time correlation for the Problem Behaviors scale was .65 (see Gresham, & Elliot, 1990, p. 111).

Interrater reliability

The SSRS manual presents correlations between mother and teacher ratings of Social Skills scale and subscale scores for a subsample of preschoolers included in the national tryout sample. These correlations ranged from .17 to .25 (all statistically significant). No correlations were presented for Problem Behaviors at the preschool level. Correlations between parent and teacher ratings for the elementary school level were also presented, based on a subsample of the national standardization sample. Correlations for Social Skills ranged from .26 to .33, and correlations for Problem Behaviors ranged from .27 to .41[11] (see Gresham, & Elliot, 1990, p. 136-137).

Validity Information from the Manual

Convergent and discriminant validity

Gresham and Elliott (1990) report results from studies examining associations between SSRS Teacher Reports and three other teacher-report measures of social skills, behavior problems, and social adjustment—the Social Behavior Assessment (SBA; Stephens, 1981), the Teacher Report Form of the Child Behavior Checklist (CBCL; Achenbach, 1991), and the Harter Teacher Rating Scale (TRS; Harter, 1985). All three studies examine ratings of elementary school children (see Gresham, & Elliot, 1990, p. 115).

- Two of the measures used in these studies—the SBA and the TRS—do not have scales or subscales that directly map onto the SSRS scales and subscales, but scores on both were expected by Gresham and Elliott (1990) to correlate with SSRS scale and subscale scores. Correlations of the SSRS scales and subscales with subscale and total scale scores on the SBA ranged from .47 to .72,[12] with one exception; SSRS Internalizing scores were correlated .19 with the SBA total score. Correlations of the SSRS scales and subscales with scores on the TRS were of similar magnitude, ranging from .44 to .70.
- The CBCL and the SSRS do have two subscales in common: Externalizing and Internalizing Problem Behaviors, as well as substantial overlap in the Total Problem Behaviors (SSRS) and Behavior Problems (CBCL) scales. Based on data from a subsample of teacher reports from the standardization sample, the correlation between total behavior problems scales from the two measures was .81. Correlations between externalizing subscales on the two measures, or between internalizing subscales on the two measures, were substantially higher than were correlations across externalizing and internalizing subscales.
    - The SSRS Externalizing subscale was correlated .75 with the CBCL Externalizing subscale, while the correlation with the CBCL Internalizing subscale was .11.
    - The SSRS Internalizing subscale was correlated .59 with the CBCL Internalizing subscale, and .19 with the CBCL Externalizing subscale.

A small subsample of parents (46) participating in the national standardization sample for the SSRS also completed the CBCL for their elementary school-age children. The Problem Behaviors scale from the SSRS correlated .70 with the CBCL Behavior Problems scale. Unlike the teacher report forms, correlations between internalizing subscales or between externalizing

[11] Gresham and Elliott include these correlations as evidence of convergent validity, rather than interrater reliability.
[12] Absolute values of correlations are presented. High scores on the SBA indicate behavior problems, while high scores on the TRS indicate positive functioning. All correlations of SSRS Total scale scores and subscales with SBA total scores and subscale scores, and with TRS scores, were in the expected direction.

subscales of the two measures were not consistently larger than were correlations across internalizing and externalizing subscales (see Gresham, & Elliot, 1990, p. 116).

- SSRS Externalizing and Internalizing subscales correlated approximately equally with the CBCL Internalizing subscale (correlations of .55 and .50, respectively).
- SSRS Externalizing and Internalizing subscales correlated .70 and .42, respectively, with the CBCL Externalizing subscale.

No similar convergent or discriminant validity studies with the preschool versions of the SSRS were included in the manual. Gresham and Elliott did include correlations between scales and subscales across parent and teacher reports as evidence of convergent and discriminant validity of the Social Skills scale and subscales for both elementary and preschool forms, and for Problem Behaviors scale and subscales for the elementary form. Information on correlations between parent- and teacher- reports of matched scales was presented in the earlier discussion of interrater reliability. Correlations between matched pairs of scales (e.g. parent-report Cooperation correlated with teacher-report Cooperation) did not differ substantially from correlations between differing subscales (e.g. parent-report Cooperation with teacher-report Self-Control) for either the preschool or elementary version.

Reliability/Validity Information from Other Studies

Internal consistency

Teacher and parent report alphas similar to those reported by Gresham and Elliott (1990) were reported for Social Skills scale scores on the elementary school version in a study of rural, low-income, white children, assessed in the fall and spring of kindergarten and in the spring of their first and second grade years, and for the Problem Behaviors scale administered during the second grade year (Pedersen, Worrell, & French, 2001).

Test-retest reliability

Pedersen, et al (2001) reported correlations between parent ratings made one year apart (from spring of kindergarten to spring of first grade, and from spring of first grade to spring of second grade) ranging from .51 to .69 for male and female children separately.

Interrater reliability

Recent studies using the current (published) version of the SSRS with samples of black children enrolled in Head Start programs have reported nonsignificant or low significant associations between parent and teacher reports (Fagan & Fantuzzo, 1999; Manz, Fantuzzo, & McDermott, 1999). Across these two studies, the strongest reported correlation, .25, was between father and teacher reports of externalizing behavior in the report by Fagan and Fantuzzo. Fagan and Fantuzzo did report more significant associations between reports from mothers and fathers. Across all families with information from both parents, six of 16 correlations were significant (r = .17 or higher), with the strongest correlations found for Internalizing and Externalizing subscales (.42 and .54, respectively). Across-parent correlations for Self-Control and Interpersonal Skills scales were lower (.34 and .17, respectively).[13]

[13] These studies utilized a different scoring system for the SSRS. Based on a series of earlier factor analyses reported by Fantuzzo, Manz, and McDermott (1998), four scales were constructed: Self-Control, Interpersonal Skills, Internalizing, and Externalizing.

In their study of low-income, rural white children, Pedersen and colleagues (2001) found correlations ranging from .53 to .57 between different teachers' ratings of children on the Social Skills scale made one year apart, but again found relatively low levels of congruence between concurrent parent and teacher ratings on the Total Social Skills scale (correlations ranging from .06 to .25).

Convergent validity

Several other researchers have independently examined associations of SSRS Social Skills and Problem Behaviors scales and subscales with scales from other assessments of children's social competence and behavior problems. Merydith (2001) correlated SSRS Social Skills and Problem Behaviors scales and Hyperactivity, Internalizing, and Externalizing Problem Behaviors subscales with parallel scales and subscales from the Behavioral Assessment System for Children (BASC; Reynolds & Kamphaus, 1998). In their ethnically mixed sample of kindergarten children, matched scales from teacher report forms of the SSRS and BASC correlated between .60 and .85. Correlations across parent report forms of the two measures ranged from .49 to .72.

Bain and Pelletier (1999) reported significant associations between teacher report SSRS Social Skills and Problem Behavior scales (particularly the Externalizing subscales) with hyperactivity and conduct problems as assessed with the Conners' Teacher Rating Scales (CTRS; Conners, 1990) in a sample of black preschoolers enrolled in Head Start programs.

Construct validity

Recent validity studies utilizing factor analyses with data from samples of black children enrolled in Head Start programs indicated subscales that did not correspond completely with those identified by Gresham and Elliott. Further, higher order factor analyses indicated that the SSRS may tap a single social competency dimension, including both Social Skills and Problem Behaviors. These findings raise questions about the discriminant validity of SSRS scales and subscales (Fantuzzo, Manz, & McDermott, 1998; Manz et al., 1999).

Comments

- The tryout version of the SSRS was modified somewhat following the national tryout, but reliability and validity information provided for the preschool version are based on the national tryout sample. Because of this, reliability and validity information presented in the manual for the preschool version may differ somewhat from the reliability and validity of the published version.
- Information provided by Gresham and Elliott (1990), as well as by Pederson and colleagues (2001) and Merydith (2001) is generally supportive of the reliability and validity of the SSRS scales and subscales. In some cases, however, information provided by Gresham and Elliott was limited to the elementary school-age form, and there were some other exceptions as well.
    - The internal consistency of parent-reported Internalizing (.57) for the preschool form was low.

- The only interrater reliability information provided by Gresham and Elliott (1990) involved correlations between ratings made by mothers and teachers. These correlations, although significant, were low to moderate. Low (and frequently nonsignificant) associations were reported by Fagan and Fantuzzo (1999) and by Manz and colleagues (1999) as well. Fagan and Fantuzzo reported somewhat higher and significant cross-parent correlations (ranging from .17 for Interpersonal Skills to .54 for Externalizing). The strongest support for interrater reliability was provided by Pederson and colleagues (2001), who reported high correlations for cross-teacher ratings conducted a year apart.
- Analyses by Gresham and Elliott (1990) provided mixed evidence of discriminant validity, particularly for parent report measures of Internalizing and Externalizing problems, as well as for the distinctiveness of the social skills subscales.
- With regard to construct validity, Gresham and Elliott present factor loadings of the SSRS items onto the factors that became the subscales for both the preschool and elementary school-age teacher- and parent-report forms. However, the manual does not provide sufficient information to allow an independent judgment of the extent to which these analyses supported the construct validity of the SSRS subscales. As indicated previously, an additional problem was that factor analyses for the preschool version were conducted with a set of items that did not correspond completely with the final published version of the measure.

- Recent studies, particularly several with black Head Start samples, raise additional concerns regarding the validity of the scale structure of the SSRS, and a different set of scales has been proposed by Fantuzzo and colleagues (1998). It may be that these results are indicative of real differences across demographic groups. Because no full-scale standardization study was conducted on the published preschool version, a modified version of the measure that was used in the tryout study, it is also possible that there are more general issues pertaining to the replicability of the factor structure for preschoolers.

VI. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

None found.

VII. Adaptations of Measure

ECLS-K Revision

Description of Adaptation

A substantial revision of the SSRS (elementary teacher- and parent-report versions) was undertaken for use in the ECLS-K. A full report of this revision was prepared by Meisels and Atkins-Burnett (1999). The total number of items was reduced, and a number of items were substantially rewritten. The Academic Competence scale found in the elementary school version of the SSRS Teacher Report form was omitted entirely. In addition, the response options for all items were modified to include a "Not Able to Observe" option that was not given in the SSRS.

- The scales that are constructed from the adapted teacher-report measure are as follows:
  - *Externalizing Problem Behaviors.*
  - *Internalizing Problem Behaviors.*
  - *Self-Control.*
  - *Interpersonal* (Positive interpersonal skills/behaviors).
  - *Task Orientation/Approaches to Learning* (This scale, composed of entirely new items not derived from the SSRS, will be discussed separately as an Approaches to Learning measure).
- The scales that are constructed from the adapted parent-report measure are as follows:
  - *Externalizing.*
  - *Internalizing.*
  - *Self-Control.*
  - *Responsibility/Cooperation.*
  - *Social Confidence.*
  - *Approaches to Learning* (As with the teacher-report, this scale will be discussed separately as an Approaches to Learning measure).

The parent and teacher importance ratings, collected but rarely used for assessment purposes in the SSRS were dropped from the revised form.

Psychometrics of Adaptation

A large, nationally representative sample of kindergartners and first graders was used for field trials of this measure. Teacher reports were completed for a total of 1,187 fall kindergartners, 1,254 spring kindergartners, and 1,286 spring first graders. Parent reports were obtained for a total of 483 fall kindergartners, 433 spring kindergartners, and 407 spring first graders. Longitudinal assessment was available for a portion of these children (i.e., children may have been tested at two or three time points).

- *Teacher Report*: The teacher-report measure tested in the field study included 41 items. Internal consistency (alpha coefficients) were examined at each time point. Reductions in the numbers of items in most scales were made, either due to unnecessarily high internal consistency (suggesting excessive redundancy), or in order to increase internal consistency.
  - Original coefficient alphas for Internalizing ranged from .73 to .75, and were improved to .76 to .80 upon modification of the items.
  - Original coefficient alphas for Externalizing ranged from .88 to .89. None were reported for the modified version of the scale.
  - Original coefficient alphas for Self Control ranged from .89 to .90. After dropping items to reduce redundancy, they ranged from .81 to .82.
  - Original coefficient alphas for the Interpersonal skills ranged from .89 to .90. After dropping items to reduce redundancy, they ranged from .85 to .86.

  Correlations of scale scores obtained in fall and spring of kindergarten ranged from .63 to .78, indicating substantial consistency in teachers' ratings of students across the school year.

Some construct validity information was provided by Meisels and Atkins-Burnett (1999) for the full 41-item measure (i.e., prior to item deletions noted above). Goodness of fit indices from confirmatory factor analyses ranged from .96 to .98, indicating a very good fit of the 5-factor model (including the Task Orientation/Approaches to Learning scale) to the data at all three time points. Despite this, however, the Self-Control scale was highly correlated with both the Interpersonal and Externalizing scales at all three waves of data collection (correlations ranging from .78 to .86). Internalizing demonstrated low to moderate correlations with the other three scales (correlations ranging from .25 to .41), and correlations between Interpersonal and Externalizing scales were moderate (.49 to .59). Thus, evidence for discriminant validity of these scales is mixed.

- *Parent Report*: The parent-report measure tested in the field study included 42 items. While exploratory factor analyses of teacher-report data indicated consistent factors across the three waves of data collection, there were differences in the numbers and content of parent-report factors at the different ages. The researchers then restricted analyses to six-factor solutions at each age and proposed retaining items that loaded consistently on the same factor at each age for the six resulting scales (including Approaches to Learning). In some cases, additional items were retained for the kindergarten grade level only. As is evident from the information provided above, alphas were frequently lower than those found for teacher-report, and indicated minimal internal consistency for some of these scales at some ages.

  - Coefficient alphas for Externalizing ranged from .67 to .72.
  - Coefficient alphas for Internalizing ranged from .55 to .59.
  - Coefficient alphas for Self-Control ranged from .50 to .70.
  - Coefficient alphas for Responsibility/ Cooperation ranged from .64 to .70.
  - Coefficient alphas for Social Confidence ranged from .66 to .74.

  Correlations of scale scores obtained in fall and spring of kindergarten ranged from .54 to .61, indicating lower consistency in parents' ratings of their children across time, compared to teacher ratings.

  Goodness of fit indices from confirmatory factor analyses of the proposed scales ranged from .90 to .99, indicating a good fit of the 5-factor model (including the Task Orientation/Approaches to Learning scale) to the data at all three time points. Correlations between all factors (excluding Approaches to Learning) ranged from .04 to .64, providing some evidence for discriminant validity of these scales.

Study Using Adaptation

- ECLS-K

Early Childhood Measures: Social-Emotional

Early Childhood Measures: Social-Emotional

Vineland Social-Emotional Early Childhood Scales (SEEC)

I. Background Information

Author/Source

Source: Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1998). *Vineland Social-Emotional Early Childhood Scales: Manual*. Circle Pines, MN: American Guidance Service, Inc.

Publisher: American Guidance Service, Inc. (AGS)
4201 Woodland Rd.
Circle Pines, MN 55014-1797
Phone: 800-328-2560
Website: www.agsnet.com

Purpose of Measure

As described by instrument publisher:
The SEEC is an assessment of social and emotional functioning in infants and young children that can be used in educational settings as well as in clinical settings to monitor individual development, to aid in the early detection of developmental delays, to help in the design of individual intervention plans, and to assess treatment effects. The manual (Sparrow, Balla, & Cicchetti, 1998) also suggests the use of the SEEC in basic research on early socioemotional development.

Population Measure Developed With

- The SEEC is derived from the Socialization Domain of the Vineland Adaptive Behavior Scales, Expanded Form (Vineland ABS; Sparrow, Balla, & Cicchetti, 1984), and norming data for the SEEC were derived from the data collected for the Vineland ABS in the early 1980s. The full norming sample included 3,000 children and adolescents ranging in age from newborns to 18 years, 11 months.
- Because the SEEC is designed specifically for young children, a subsample of 1,200 children ranging in age from newborn through 5 years, 11 months, constituted the SEEC norming sample. This subsample included 100 children representing each of 12 6-month age groups from newborns to age 6 who were selected to create a sample that resembled, as closely as possible, the U.S. population within the age range with respect to ethnicity, gender, community size, geographic region, and parent education, according to 1980 U.S. Census data.
- Approximately half (49.4 percent) of the sample were females, half (50.6 percent) males. Parent education was distributed similarly to the U.S. population, although somewhat fewer parents in the sample had less than high school education than did adults ages 20 through 44 in the population (12.8 percent vs. 15.7 percent). White children were slightly overrepresented, and Hispanic children slightly underrepresented, relative to the U.S. population (75.7 percent vs. 72.0 percent for white children, 7.6 percent vs. 10.1 percent

for Hispanic children). Slightly more than half (58 percent) of the 3- to 5-year-olds in the sample were enrolled in preschool or school programs.

Age Range Intended For

Newborn through age 5 years, 11 months.

Key Constructs of Measure

The SEEC is composed of three subscales that combine to form a *Social-Emotional Composite* score. Each subscale includes items that are ordered according to the ages at which children would be expected to achieve the behavior described, beginning with items that are appropriate in very early infancy and moving through items that should be attained through the preschool years. All items were included in the original ABS, and the subscales are the same as those obtained for the Socialization Domain of the ABS. All Socialization Domain items from the ABS that were appropriate for young children and that were reported to have been exhibited by at least one percent of children younger than age 6 in the standardization sample were retained for the SEEC. A total of 12 ABS items were dropped. The three subscales include:

- *Interpersonal Relationships*: Responsiveness to social stimuli, age-appropriate emotional expression and emotion understanding, age-appropriate social interactive behaviors, ability to make and maintain friendships, and cooperation.
- *Play and Leisure Time:* Characteristics of toy play, interest in the environment and exploration, social play, make-believe activities, interest in television, playing games and following rules, hobbies and activities, independence.
- *Coping Skills:* This subscale is not administered to toddlers and infants under 2 years of age. Compliance with rules, politeness, responsibility, sensitivity to others, impulse control.

Norming of Measure (Criterion or Norm Referenced)

Norm referenced.

Comments

- Behavior problems are not directly assessed with this measure—only relative strengths and weaknesses in adaptive functioning.
- Although based on a measure that has been in use for nearly 20 years, the SEEC is a very new measure and its usefulness as a free-standing measure is not well established at this time.
- The establishment of "new" norms using data that are nearly 20 years old is unusual. The validity of doing this assumes that the items and scales on the measure have not themselves become antiquated due to changing cultural norms, and that demographic shifts in the U.S. population will not dramatically affect the population norms for the items and scales. Examination of the items and scales of the SEEC do not suggest that these assumptions are incorrect, however some caution may be warranted in using normed scores on this measure.

II. Administration of Measure

Who is the Respondent to the Measure?
Parent, guardian, or adult caregiver.

If Child is Respondent, What is Child Asked to Do?
N/A.

Who Administers Measure/Training Required?
Test Administration:

- Sparrow and colleagues (1998) indicate that interviewers should have education and experience pertaining to child development and behavior, as well as in tests and measurement, and should be trained in interview techniques. Further, interviewers should have specific training (which can be accomplished through practice sessions) in administering and interpreting the Vineland SEEC.

Data Interpretation:

- (Same as above.)

Setting (e.g., one-on-one, group, etc.)
One-on-one.

Time Needed and Cost
Time:

- The SEEC takes 15-25 minutes to administer.

Cost:

- Vineland SEEC Kit (including manual): $57.95
- Record forms: $26.95 per package of 25 forms

Comments

- The SEEC is one of the few measures of social–emotional development that can be used with children from birth onward.
- Administration of the SEEC is very different from other social-emotional measures, most of which can be easily administered as questionnaires. It requires a one-on-one interview, conducted by a highly trained interviewer who has had a great deal of practice with the measure. Interviews are largely unstructured and interviewers must attain a high level of competence in conducting interviews from which the necessary information can be obtained without leading the parent to respond in particular ways, or overlooking important information entirely. Thus, administration of the SEEC would be more costly to administer than other social-emotional measures. It is unclear whether there would be benefits to the SEEC that outweigh these higher costs.

III. Functioning of Measure

Reliability Information from the Manual

Internal consistency

Internal consistencies of the subscales and the total Social Emotional Composite were established in the norming sample for each 1-year age level (e.g., 0 years through 11 months, 1 year through 1 year, 11 months) using a modified split-half procedure. Median split-half reliability coefficients were .84 for Interpersonal Relationships, .80 for Play and Leisure Time, .87 for Coping Skills, and .93 for the Social Emotional Composite (see Sparrow et al., 1998, p. 85).

Test-retest reliability

A subsample of 182 children from the norming sample in the age range covered by the SEEC (70 children ages 6 months through 2 years, 11 months and 112 children ages 3 years through 5 years, 11 months) were assessed twice by the same interviewer, with a mean interval of 17 days (ranging from 2 to 4 weeks) between parent interviews. Test-retest correlations for all subscales and the Social Emotional Composite ranged from .71 to .79, with one exception; the test-retest correlation for Coping Skills in the younger age group (ages 2 years through 2 years, 11 months, $N = 33$) was .54. The authors recommend caution in interpreting scores on this subscale for younger children or children with low abilities (see Sparrow et al., 1998, p. 87).

A smaller subsample of 78 children from the norming sample ages 6 months through 5 years, 11 months were assessed twice by two different interviewers, with an approximately 1-week interval between parent interviews. Test-retest correlations were lower when assessments were conducted by different interviewers, ranging from .47 to .60 (see Sparrow et al., 1998, p. 88).

Validity Information from the Manual

Sparrow and colleagues indicate that, because one method of interpreting scores on the SEEC is in terms of age equivalents, it is important for the validity of the measure that total raw scores increase systematically across age. Data from the norming sample indicated that expected age increases did occur with no backward steps, although the increases became progressively smaller at the older ages (see Sparrow et al., 1998, p. 90). For example, the mean Interpersonal Relationships score for children ages newborn to 5 months was 16.3 ($SD = 6.5$), while the mean score for children ages 6 months through 11 months was 28.1 ($SD = 5.6$), a change of nearly 12 points. Between the ages of 5 years through 5 years, 5 months and 5 years, 6 months through 5 years, 11 months, in contrast, there was only a 1-point increase, from 70.5 ($SD = 6.4$) to 71.5 ($SD = 7.5$).

Construct validity

Sparrow and colleagues present correlations among the three subscales separately for each 1-year age group of the norming sample. Correlations between the three subscales (two subscales for the two youngest groups) were similar at all ages, ranging from .45 to .66 (see Sparrow et al., 1998, p. 91).

Convergent validity

Sparrow and colleagues (1998) cite independent studies that have found significant correlations between the Vineland ABS Socialization Domain scores and scores from other measures tapping similar constructs. Specifically, correlations ranging from .51 to .65 have been found between infants' and young children's ABS Socialization Domain scores and Personal-Social Domain scores from the Battelle Developmental Inventory (Johnson, Cook, & Kullman, 1992) and scores on the long and short forms of the Scales of Independent Behavior (Goldstein, Smith, Waldrep, Inderbitzen, 1987).

Discriminant validity

A portion of the norming sample for the SEEC (222 children, ages 2 years, 6 months through 5 years, 11 months) was also included in the norming sample for the Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983), an assessment of cognitive ability and achievement. Correlations were expected by Sparrow and colleagues to be low, because these are measures tapping different domains of functioning. Reported correlations between SEEC composites and K-ABC measures ranged from .02 to .18. Similarly, correlations between SEEC subscale and composite scores and scores on the Peabody Picture Vocabulary Test – Revised (PPVT-R; Dunn & Dunn, 1981) ranged from .15 to .19 in a subsample of 559 children ages 2 years, 11 months through 5 years, 11 months who were part of the SEEC norming sample (see Sparrow et al., 1998, p. 94-95).

Concurrent and predictive validity

Sparrow and colleagues (1998) present a brief review of studies that have found associations between Vineland ABS Socialization Domain scores and characteristics of children, such as autism (e.g., Vig & Jedrysek, 1995; Volkmar, Sparrow, Goudreau, & Cicchetti, 1987), and of their environments, such as maternal child-rearing practices (Altman & Mills, 1990) that would be expected to affect social-emotional functioning. No studies with the newly constructed SEEC measure were reported, however.

Reliability/Validity Information from Other Studies

- None found.

Comments

- Overall, information provided by Sparrow and colleagues regarding reliability of the SEEC scales suggests that the scales have high levels of internal consistency and high test-retest reliability when administered by the same tester (with the exception of Coping Skills at ages below 3 years). Test-retest correlations were lower (although still moderate to high) when the SEEC was administered by different interviewers, which may indicate that characteristics of the interviewer have some effects on information obtained in the interview process (including how parents report on children's behavior), and/or on how the interview content is interpreted and summarized. At a minimum it underscores the necessity for careful initial interviewer training and frequent retraining in order to avoid drift in assessment practices.
- With respect to validity, moderate to high correlations presented by Sparrow and colleagues (1998) between the three SEEC scales (Interpersonal Relationships, Play and Leisure Time, and for children age 2 and older, Coping Skills) provide some support for the validity of these scales as measures of distinctive yet interrelated aspects of social-emotional

functioning in infancy and early childhood. The authors also present clear support for the convergent and discriminant validity of SEEC scales; correlations with other measures of social-emotional functioning were consistently high, while correlations with measures of cognitive functioning were consistently low.

- Some caution may be called for when interpreting these findings, however. All information on reliability and validity of the SEEC is based on data from a sample of individuals who were actually interviewed more than a decade earlier with a longer instrument, the Vineland ABS, and the abbreviated measure was developed based on statistical results involving this longer measure. Because the properties of scales can be affected by the context in which they are presented (including other items and scales included in the measure), the information provided by Sparrow and colleagues (1998) regarding the SEEC as a separate measure (administered apart from the other items on the Vineland ABS) may not fully reflect properties of the newly-constructed measure.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- No studies with the Vineland SEEC were found. One fairly recent study did report findings involving the Interpersonal Relationships subdomain of the Vineland ABS (part of the Socialization Domain of the ABS, from which the SEEC was derived). Marcon (1999) found that 4-year-old black children attending preschool programs described as "child initiated" (i.e., programs in which children's learning activities are self-directed, facilitated and encouraged but not controlled by teachers and where little emphasis is placed on direct instruction) had higher scores on Vineland ABS Interpersonal Relationships subdomain scores than did children enrolled in "adult directed" preschool programs (i.e., more academically-oriented programs focusing on direct instruction and teacher-directed learning experiences).

V. Adaptations of Measure

None found.

Early Childhood Measures:
Approaches to Learning

Early Childhood Measures: Approaches to Learning

ECLS-K[14] Adaptation of the Social Skills Rating System (SSRS), Task Orientation/Approaches to Learning Scale

I. Background Information

Author/Source

Source: Meisels, S. J., & Atkins-Burnett, S. (1999). *Social Skills Rating System field trial analysis report and recommendations*. Final project report prepared for National Opinion Research Center.

Publisher: Documentation available from the National Center for Education Statistics (NCES).

Purpose of Measure

As described by developer:
This measure is an assessment of task orientation, adaptability, motivation, and creativity. It is a component of a longer measure tapping social skills and problem behaviors, a revision of the Social Skills Rating System (SSRS; Gresham & Elliott, 1990), designed by Meisels and Atkins-Burnett (1999) for use in the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K). Approaches to learning are not specifically assessed in the SSRS, and items on this scale are new or have been substantially modified.

Population Measure Developed With

- A large, nationally-representative sample of kindergartners and first graders was used for field trials of this measure.
- Teacher reports were completed for a total of 1,187 Fall kindergartners, 1,254 Spring kindergartners, and 1,286 Spring first graders.
- Parent reports were obtained for a total of 483 Fall kindergartners, 433 Spring kindergartners, and 407 Spring first graders. Longitudinal assessment was available for a portion of these children (i.e., children may have been tested at two or three time points).

Age Range Intended For

Kindergartners and first graders.

Key Constructs of Measure

This 6-item scale primarily assesses behaviors related to engagement in learning, organization, creativity, and adaptability.

Norming of Measure (Criterion or Norm Referenced)

No norming has been done with this scale. Extensive information on means and standard deviations for the sample used for field testing, and for subsamples broken down by ethnicity, gender, and other demographic characteristics are included in the documentation.

[14] Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999.

Comments

- This scale has not been used with pre-kindergartners. With one possible exception ("Keeps belongings organized"), however, the six items included in the ECLS-K appear to be developmentally appropriate for younger preschoolers.
- An additional concern is that there is no overlap in item content across the parent- and teacher-report forms. The overall concepts covered by the two scales appear to be similar; however the parent-report form includes two items that appear to tap behaviors that are not reflected in the teacher-report—an item pertaining to creativity and an item involving helping with chores. The latter of these items may be more reflective of child compliance than approaches to learning.

II. Administration of Measure

Who is the Respondent to the Measure?

Teacher and parent.

If Child is Respondent, What is Child Asked to Do?

N/A.

Who Administers Measure/Training Required?

Test Administration:

- In the ECLS-K, these items were administered as part of a larger survey by experienced field staff. No particular training should be required to administer these short measures.

Setting (e.g., one-on-one, group, etc.)

One-on-one or independent. Teachers and parents generally complete these brief measures either on their own or by responding aloud to questions posed as part of a survey in a one-to-one setting. In ECLS-K field trials, parent survey administration was done via computer-assisted telephone interviews (CATI) while teachers completed a paper-and-pencil version.

Time Needed and Cost

Time:

- Teacher report administration time is very brief. Administration of the full ECLS-K adaptation of the SSRS takes approximately 5 to 6 minutes per child.

III. Functioning of Measure

Reliability

Internal Consistency

The coefficient alpha for the teacher-report version of this scale in the field trials was .89.

Documentation regarding the parent-report version of this scale in the field trials is somewhat unclear. The measure included either four or five items at that time. The alpha coefficient for

this scale ranged from .72 to .77 at the three assessments points (Fall of kindergarten, Spring of kindergarten, Spring of 1st grade).

Test-Retest Reliability
Teacher ratings of children's Task Orientation/Approaches to Learning during the Fall and Spring of kindergarten were found to correlate .77.

Parent ratings of children's Approaches to Learning across the same interval correlated .55.

Interrater Reliability
Correlations between parent and teacher reports of children's Approaches to Learning ranged from .16 to .19 at the three assessment points.

Validity

In the ECLS-K Field Test, correlations between teacher-report Approaches to Learning and teacher ratings of academic performance in language and literacy, math, and general knowledge in kindergarten and first grade ranged from .51 to .66. Correlations with direct cognitive test scores for reading, general knowledge, and math were significant although somewhat lower, ranging from .31 to .47. Correlations between kindergartners' Fall Approaches to Learning teacher ratings and measures of gains in reading, general knowledge, and math from Fall to Spring were generally nonsignificant, however.

Parent reports of kindergartners' Approaches to Learning were also significantly correlated with teacher ratings of academic performance in language and literacy, math, and general knowledge, although the correlations were substantially lower, ranging from .16 to .24. Parent-reported Approaches to Learning in the fall was also correlated significantly with direct assessments of kindergartners' reading and math scores (but not general knowledge) conducted the following spring. Significant correlations were also found between Fall Approaches to Learning parent ratings and children's gains in reading achievement from Fall to Spring (correlations of .18 to .21).

Reliability/Validity Information from Other Studies

None found.

Comments

- The information available regarding the functioning of the ECLS-K Approaches to Learning measures is considerably clearer for teacher-report than for parent-report, which underwent substantial revision following field testing.
- Overall, these short measures appear to have strong internal consistency and high test-retest reliability across an interval of several months.
- Associations between concurrent parent and teacher reports were low. However, this should not be surprising given that the construction of the parent and teacher report scales differs substantially and given the very different contexts in which parents and teachers are likely to most frequently observe the child's learning activities.

- The validity of these measures was assessed by examining associations between concurrent and subsequent academic performance. These analyses generally indicated expected positive associations, although concurrent associations were considerably stronger than predictive associations, and expected associations between teachers' ratings of children's Approaches to Learning early in the school year and gains in reading, math, and general knowledge from fall to spring were nonsignificant. Interestingly, there were significant low correlations between measures of academic gains and parent-reported Approaches to Learning.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

None found.

V. Adaptations of Measure

None found.

References for Social-Emotional and Approaches to Learning Measures

Achenbach, T.M. (1991). *Manual for the Child Behavior Checklist/4-18*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T.M. (1992). *Manual for the Child Behavior Checklist/2-3 and 1992 profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T.M. (1997). *Guide for the Caregiver-Teacher Report Form for Ages 2-5.* Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T.M., Edelbrock, C., & Howell, C.T. (1987). Empirically based assessment of the behavioral/emotional problems of 2- and 3-year-old children. *Journal of Abnormal Child Psychology, 15*, 629-650.

Achenbach, T.M., Howell, C., Aoki, M., & Rauh, V. (1993). Nine-year outcome of the Vermont Intervention Program for Low Birth Weight Infants. *Pediatrics, 91*(1), 45-55.

Achenbach, T.M., & Rescorla, L.A. (2000). *Manual for the ASEBA preschool forms & profiles.* Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Achenbach, T.M., & Rescorla, L.A. (2001). *Manual for the ASEBA school-age forms & profiles.* Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Altman, J.S., & Mills, B.C. (1990). Caregiver behavior and adaptive behavior development of very young children in home care and daycare. *Early Child Development and Care, 62*, 87-96.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Association.

Arend, R., Lavigne, J.V., Rosenbaum, D., Binns, H.J., & Christoffel, K.K. (1996). Relation between taxonomic and quantitative diagnostic systems in preschool children: Emphasis on disruptive disorders. *Journal of Clinical Child Psychology, 25*(4), 388-397.

Bain, S.K., & Pelletier, K.A. (1999). Social and behavioral differences among a predominantly African American preschool sample. *Psychology in the Schools, 36*(3), 249-259.

Baker, P.C., Keck, C.K., Mott, F.L., & Quinlan, S.V. (1993). *NLSY child handbook, revised edition: A guide to the 1986-1990 National Longitudinal Survey of Youth Child Data*. Columbus, OH: Center for Human Resource Research, The Ohio State University.

Bayley, N. (1969). *Manual for the Bayley Scales of Infant Development*. New York: Psychological Corporation.

Briggs-Gowan, M.J., & Carter, A.S. (1998). Preliminary acceptability and psychometrics of the Infant-Toddler Social and Emotional Assessment (ITSEA): A new adult-report questionnaire. *Infant Mental Health Journal, 19*(4), 422-445.

Briggs-Gowan, M.J., Carter, A.S., Skuban, E., & McCue Horwitz, S. (2001). Prevalence of social-emotional and behavioral problems in a community sample of 1- and 2-year-old children. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*(7), 811-819.

Brown, L.L., & Hammill, D.D. (1983). *Behavior Rating Profile (BRP-2)*. Austin, TX: PRO-ED, Inc.

Burks, H.F. (1977). Burks' Behavior Rating Scales (BBRS). Austin, TX: PRO-ED, Inc.

Capuano, F., & LaFreniere, P.J. (1993). *Early identification and prevention of affective disorders in children.* Paper presented at the National Head Start Conference, Washington, DC.

Cohen, N.J. (1983). Mother-child interaction in hyperactive and normal kindergarten-aged children and the effect of treatment. *Child Psychiatry & Human Development, 13*(4), 213-224.

Conners, C.K. (1989). *Conners' Parent Rating Scales*. North Tonawanda, NY: Multi-Health Systems.

Conners, C.K. (1990). *Manual for Conners' Rating Scales*. Toronto: Multi-Health Systems Inc.

Conners, C.K. (1995). *Conners' Continuous Performance Test (CPT)*. Toronto: Multi Health Systems Inc.

Conners, C.K. (1997). *Conners' Rating Scales – Revised: Technical manual*. North Tonawanda, NY: Multi-Health Systems Inc.

Dubow, T. & Luster, E. (1990). Predictors of the quality of the home environment that adolescent mothers provide for their school-aged children. *Journal of Marriage & the Family*, 52(2), 393-404.

Dumas, J.E., & LaFreniere, P.J. (1993). Mother-child relationships as sources of support or stress: A comparison of competent, average, and anxious dyads. *Child Development, 64*, 1732-1754.

Dumas, J. E., Martinez, A., & LaFreniere, P. J. (1998). The Spanish version of the Social Competence and Behavior Evaluation (SCBE)--Preschool Edition: Translation and field testing. *Hispanic Journal of Behavioral Sciences, 20*(2), 255-269.

Dunn, L.M., & Dunn, L.M. (1997). *Peabody Picture Vocabulary Test: Third edition*. Circle Pines, MN: American Guidance Service, Inc.

Elliott, S. N., Barnard, J., & Gresham, F. M. (1989). Preschoolers' social behavior: Teachers' and parents' assessments. *Journal of Psychoeducational Assessment, 7*(3), 223-234.

Fagan, J., & Fantuzzo, J. W. (1999). Multirater congruence on the Social Skills Rating System: Mother, father, and teacher assessments of urban Head Start children's social competencies. *Early Childhood Research Quarterly, 14*(2), 229-242.

Fantuzzo, J., Manz, P. H., & McDermott, P. (1998). Preschool version of the Social Skills Rating system: An empirical analysis of its use with low-income children. *Journal of School Psychology, 36*(2), 199-214.

Flanagan, D.P., Alfonso, V.C., Primavera, L.H., Povall, L., & Higgins, D. (1996). Convergent validity of the BASC and SSRS: Implications for social skills assessment. *Psychology in the Schools, 33*, 13-23.

Flanagan, R. (1995). A review of the Behavior Assessment System for Children (BASC): Assessment consistent with the requirements of the Individuals with Disabilities Education Act (IDEA). *Journal of School Psychology, 33*(2), 177-186.

Garmezy, N. (1985). Stress-resistant children: the search for protective factors. In J.E. Stevenson (Ed.), *Recent research in developmental psychopathology. Journal of Child Psychology and Psychiatry* (Book Supplement, No. 4, pp. 213-233). Oxford: Pergamon Press.

Gennetian, L.A., & Miller, C. (2002). Children and welfare reform: A view from an experimental welfare program in Minnesota. *Child Development, 73*(2), 601-620.

Gilliom, M., Shaw, D.S., Beck, J.E., Schonberg, M.A., & Lukon, J.L. (2002). Anger regulation in disadvantaged preschool boys: Strategies, antecedents, and the development of self-control. *Developmental Psychology, 38*(2), 222-235.

Goldstein, D.J., Smith, K.B., Waldrep, E., & Inderbitzen, H.M (1987). Comparison of the Woodcock-Johnson Scales of Independent Behavior and Vineland Adaptive Behavior Scales in infant assessment. *Journal of Psychoeducational Assessment, 5*(1), 1-6.

Gresham, F.M., & Elliott, S.N. (1990). Social Skills Rating System: Manual. Circle Pines, MN: American Guidance Service.

Harter, S. (1985). *Manual for the Self-Perception Profile for Children*. University of Denver, Denver, CO.

Hogan, A.E., Quay, H.C., Vaughn, S., & Shapiro, S.K. (1989). Revised Behavior Problem Checklist: Stability, prevalence, and incidence of behaivor problems in kindergarten and first grade children. *Psychological Assessment, 1*(2), 103-111.

Ireton, H., & Thwing, E. (1974). *Minnesota Child Development Inventory.* Minneapolis, MN: Behavior Science Systems.

Johnson, L.J., Cook, M.J., & Kullman, A.J. (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*(4), 353-359.

Kaiser, A.P., Hancock, T.B., Cai, X., Foster, E.M., & Hester, P.P. (2000). Parent-reported behavioral problems and language delays in boys and girls enrolled in head start classrooms. *Behavioral Disorders, 26*(1), 26-41.

Kaufman, A.S., & Kaufman, N.L. (1983). *Kaufman Assessment Battery for Children: Administration and scoring manual*. Circle Pines, MN: American Guidance Service, Inc.

Kaufman, A.S., & Kaufman, N.L. (1983). *Kaufman Assessment Battery for Children: Interpretive manual*. Circle Pines, MN: American Guidance Service, Inc.

Keenan, K., & Wakschlag, L.S. (2000). More than the terrible twos: The nature and severity of behavior problems in clinic-referred preschool children. *Journal of Abnormal Child Psychology, 28*(1), 33-46.

Koot, H., van den Oord, J., Verhulst, F.C., & Boomsma, D. (1997). Behavioral and emotional problems in young preschoolers: Cross-cultural testing of the validity of the Child Behavior Checklist/2-3. *Journal of Abnormal Child Psychology, 25*, 183-196.

Kovacs, M. (1992). *Children's Depression Inventory (CDI): Manual*. Toronto: Multi-Health Systems Inc.

Lacher, D. (1982). *Personality Inventory for Children – Revised*. Los Angeles: Western Psychological Services.

LaFreniere, P.J., & Capuano, F. (1997). Preventive intervention as means of clarifying direction of effects in socialization: Anxious-withdrawn preschoolers case. *Development & Psychopathology, 9*(3), 551-564.

LaFreniere, P.J., & Dumas, J.E. (1992). A transactional analysis of early childhood anxiety and social withdrawal. *Development and Psychopathology, 4*, 385-402.

LaFreniere, P.J., & Dumas, J.E. (1995). *Social Competence and Behavior Evaluation—Preschool Edition (SCBE).* Los Angeles: Western Psychological Services.

LaFreniere, P.J., & Dumas, J.E. (1996). Social competence and behavior evaluation in children ages 3 to 6 years: The short form (SCBE-30). *Psychological Assessment, 8*(4), 369-377.

LaFreniere, P.J., Dumas, J.E., Capuano, F., & Dubeau, D. (1992). Development and validation of the Preschool Socio-Affective Profile. *Psychological Assessment, 4*(4), 442-450.

LaFreniere, P., Masataka, N., Butovskaya, M., Chen, Q., Dessen, M.A., Atwanger, K., et al. (2002). Cross-cultural analysis of social competence and behavior problems in preschoolers. *Early Education & Development, 13*(2), 201-219.

LeBuffe, P.A., & Naglieri, J.A. (1999). *Devereux Early Childhood Assessment Program: Technical manual.* Lewisville, NC: Kaplan Press.

LeBuffe, P.A., & Naglieri, J.A. (n.d.). *The Devereux Early Childhood Assessment (DECA). A measure of within-child protective factors in preschool children.* Retreived June 9, 2002, from http://www.devereuxearlychildhood.org/DECI/monograph.htm

Manz, P. H., Fantuzzo, J. W., & McDermott, P. A. (1999). The parent version of the Preschool Social Skills Rating Scale: An analysis of its use with low-income, ethnic minority children. *School Psychology Review, 28*, 493-504.

Marcon, R.A. (1999). Differential impact of preschool models on development and early learning of inner-city children: A three-cohort study. *Developmental Psychology, 35*(2), 358-375.

McCarthy, D. (1972). *McCarthy Scales of Children's Abilities*. New York: Psychological Corporation.

Meisels, S.J., & Atkins-Burnett, S. (1999). *Social Skills Rating System field trial analysis report and recommendations*. Final project report prepared for National Opinion Research Center.

Merydith, S.P. (2001). Temporal stability and convergent validity of the Behavior Assessment System for Children. *Journal of School Psychology, 39*(3), 253-265.

Mouton-Simien, P., McCain, A.P., & Kelley, M.L. (1997). The development of the Toddler Behavior Screening Inventory. *Journal of Abnormal Child Psychology, 25*, 59-64.

National Center for Health Statistics (1982). Current estimates from the National Health Interview Survey: United States, 1981. Public Health Service, *Vital and Health Statistics,* Series 10, No. 141. DHHS Pub. No. (PHS) 83-1569. Washington, DC: U.S. Government Printing Office.

Pedersen, J. A., Worrell, F. C., & French, J. L. (2001). Reliability of the Social Skills Rating System with rural Appalachian children from families with low incomes. *Journal of Psychoeducational Assessment, 19*, 45-53.

Pelham, W.E., Swanson, J.M., Furman, M.B., & Schwindt, H. (1996). Pemoline effects on children with ADHD: A time-response by dose-response analysis on classroom measures. *Annual Progress in Child Psychiatry & Child Development*, 473-493.

Peterson, J. L., & Zill, N. (1986). Marital disruption, parent-child relationships, and behavior problems in children. *Journal of Marriage and the Family, 48*, 295-307.

Reynolds, C.R., & Kamphaus, R.W. (1998). *BASC Behavioral Assessment System for Children: Manual.* Circle Pines, MN: American Guidance Service, Inc.

Richman, N., Stevenson, J., & Graham, P.J. (1982). *Pre-school to school: A behavioural study*. London and New York: Academic Press.

Rust, L.W. (2001). *Summative evaluation of Dragon Tales. Final report.* Retreived from http://pbskids.org/dragontales/caregivers/about/dt_eval_final_report.pdf

Shaw, D.S., Keenan, K., Vondra, J.I., Delliquadri, E., & Giovannelli, J. (1997). Antecedents of preschool children's internalizing problems: A longitudinal study of low-income families. *Journal of the American Academy of Child & Adolescent Psychiatry, 36*(12), 1760-1767.

Sparrow, S.S., Balla, D.A., & Cicchetti, D.V. (1984). *Vineland Adaptive Behavior Scales.* Circle Pines, MN: American Guidance Service, Inc.

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

Spiker, D., Kraemer, H.C., Constantine, N.A., & Bryant, D. (1992). Reliability and validity of behavior problem checklists as measures of stable traits in low birth weight, premature preschoolers. *Child Development, 63*, 1481-1496.

Stephens, T. (1981). *Technical manual: Social Behavior Assessment*. Columbus, OH: Cedars Press.

Vandell, D. L.& Ramanan, J. (1991). Children of the National Longitudinal Survey of Youth: Choices in after-school care and child development. *Developmental Psychology*, 27(4), 637-643.

Vig, S., & Jedrysek, E. (1995). Adaptive behavior of young urban children with developmental disabilities. *Mental Retardation, 33*(2), 90-98.

Volkmar, F., Sparrow, S., Goudreau, D., & Cicchetti, D. (1987). Social deficits in autism: An operational approach using the Vineland Adaptive Behavior Scales. *The Journal of the American Academy of Child and Adolescent Psychiatry, 26*(2), 151-161.

Werner, E.E., & Smith, R.S. (1982). *Vulnerable but invincible: A longitudinal study of resilient children*. New York: Adams, Bannister, Cox.

Early Childhood Measures:
Ongoing Observational

Early Childhood Measures: Ongoing Observational

High/Scope Child Observation Record (COR)

I. Background Information

Author/Source

Source: Schweinhart, L., McNair, S., Barnes, H., & Larner, M. (1993). Observing young children in action to assess their development: The High/Scope Child Observation Record Study. *Educational and Psychological Measurement, 53,* 445-454.

Publisher: High/Scope Educational Research Foundation
600 North River St.
Ypsilanti, MI 48198
Phone: 734-485-2000
Website: www.highscope.org

Purpose of Measure

As described by instrument publisher:
"The High/Scope Child Observation Record for ages 2 ½ - 6 (COR) is an observational assessment tool that can be used in a variety of early childhood settings…It is developmentally appropriate, both in breadth of content and in process. COR assessment areas include not only language and mathematics, but also initiative, social relations, creative representation, and music and movement" (from Website; see www.highscope.org/Assessment/cor.htm).

Population Measure Developed With

- This observational assessment tool is intended as a vehicle for documenting development and progress over time. Records from the observations are not related to norms from a standardization sample.
- Measure development and psychometric work were carried out in a sample of about 2,500 children. The data come from seven Head Start agencies and one school district in southeastern Michigan. The sample was diverse; 51 percent black, 26 percent white, 14 percent Arab, 7 percent Hispanic, 2 percent Asian/Pacific Islander, and 1 percent American Indian/Alaskan Native.

Age Range Intended For

Ages 2 years, 6 months through 6 years.

Key Constructs of Measure

- This measure focuses on six constructs, each involving several skills (see Table 1 below for skills comprising each of the constructs):
  - *Initiative*
  - *Social Relations*
  - *Creative Representation*
  - *Music and Movement*

- *Language and Literacy*
- *Logic and Mathematics*

Norming of Measure (Criterion or Norm Referenced)
Criterion referenced.

Comments
- The proportion of the sample from Head Start versus public schools was not noted. Having included seven Head Start agencies, it is assumed that this observational system is appropriate for Head Start, but systematic differences between the Head Start and public school samples were not assessed.
- While the racial/ethnic distribution in the sample was different from that for Michigan in the 2000 Census, the fact that the sample included children from a range of racial/ethnic groups suggests that the measure is appropriate for children from differing backgrounds.
- It is not clear how the assessment tool would work for children with special needs. However, this issue is less salient for a measure in which children's development is primarily related to their own progress over time than for a measure that relies on norms from a standardization sample.

II. Administration of Measure

Who is the Respondent to the Measure?
Teacher.

If Child is Respondent, What is Child Asked to Do?
N/A.

Who Administers Measure/Training Required?
Administration:
- The COR is meant to work closely with methods used by schools to document children's progress (e.g., portfolios, checklists, notes, or mixtures of these). For instance, teachers might take notes on instances in which children illustrate knowledge of letters and an increased ability to write their names, or they might collect samples of the children's work that illustrates such growth. In the measure development/validation study, teachers wrote brief notes on index cards over the course of the school year describing the six aspects of development noted above for each child in their class.
- Once teachers have recorded information on individual children for a substantial period of time, they are asked to assess each child's level on a series of skills within each construct. This is done by choosing from a list of continuous indicators for each skill (e.g., Expressing choices, Solving problems, Engaging in complex play, etc.) within the larger construct (e.g., Initiative). For example, indictors for the "Expressing Choices" skill in the "Initiative" category are as follows:
  - Child does not yet express choice to others.
  - Child indicates a desired activity or place of activity by saying a word, pointing, or some other action.

- Child indicates desired activity, place of activity, materials, or playmates with a short sentence.
- Child indicates with a short sentence how plans will be carried out ("I want to drive the truck on the road").
- Child gives detailed description of intended actions ("I want to make a road out of blocks with Sarah and drive the truck on it").

- Constructs and skills included in the assessment are presented in Table 1.

Table 1.

Initiative	Social Relations	Creative Representation	Music and Movement	Language and Literacy	Logic and Mathematics
- Expressing choices - Solving problems - Engaging in complex play - Cooperating in routines	- Relating to adults - Relating to children - Making friends - Solving social problems - Expressing feelings	- Making & building - Drawing & painting - Pretending	- Body & coordination - Manual coordination - Imitating a beat - Movement & directions	- Understanding speech - Speaking - Interest in reading - Using books correctly - Beginning reading - Beginning writing	- Arranging in order - Using comparison words - Sorting - Using, some, not, & all - Comparing numbers - Counting objects - Spatial relations - Sequence & time

Training:

- In the measure development/validation study, each teacher/teaching assistant attended a three-day COR training session led by a professional trainer.
- There was also ongoing follow-up (although the extent of follow-up and consistency across sites is not clear). The project coordinator maintained ongoing quality control after the training by reviewing the anecdotal notes that teachers recorded about children, and by scheduling feedback and additional training sessions as needed. Head Start education coordinators were also involved in the ongoing process of visiting classrooms and holding training sessions. It was unclear whether there was a counterpart to the Head Start education coordinator position in the public school based part of the sample.

Data Interpretation:

- The teacher who maintains records for a child and completes the skill level ratings also interprets the results, using them to guide activities and instruction, and provide information to parents.

Setting (e.g., one-on-one, group, etc.)

The teachers observe individual children over time, but the context for observations may be a group setting.

Time Needed and Cost

Time:

- Ongoing

Cost:

- $124.95

Comments

- It is noteworthy that ongoing follow-up was built into training. However, the recommended frequency/extent of follow-up is not clear.

III. Functioning of Measure

Reliability

Internal Consistency

Internal consistency (Cronbach's alpha) for each of the six construct level scales ranged from .80 to .93 (median = .87) for teacher ratings and .72 to .91 (median = .85) for teacher assistants (see Schweinhart, McNair, Barnes, & Larner, 1993, p. 450). As internal consistency was at the construct level, each of the skills (e.g., Expressing choices, Solving problems, Engaging in complex play, etc.) within the given construct was treated as a continuous to assess coefficient alphas.

Interrater Reliability

Each teacher and teaching assistant independently completed CORs on the same 10 children. Correlations between ratings by teachers and assistants ranged from .62 to .72 (see Schweinhart, McNair, Barnes, & Larner, 1993, p. 452). In a related study (Epstein, 1992), 10 research assistants were trained to a level of agreement of .93 (kappa scoring), after three days of training.

Validity

Construct Validity

Confirmatory factor analysis was done for the six COR scales and maximum likelihood estimates were assessed for the factor loadings, ranging from .62 to .86, with two items below .70. The goodness-of-fit for the entire index was .789.

Concurrent Validity

Ninety-eight children for whom the COR was being completed were also administered the McCarthy Scales of Children's Ability (MSCA; McCarthy, 1972). Criteria for selection of this sample are unclear. Scale scores from the COR were related to scores from the McCarthy scales for General Cognition, Verbal Ability, Perceptual Performance, Quantitative, Memory, and Motor Skills.

Correlations ranged from .27 to .66, with most correlations falling in the low to moderate range. The COR Language and Literacy scale showed the greatest relation to the McCarthy scales, with

correlations ranging from .53 to .66. Correlations between the COR scale scores and MSCA scores were as follows:

- For the COR Initiative scale, correlations with MSCA scores ranged from .31 to .43 (the lowest correlation was with MSCA Verbal and the highest with Perceptual Performance).
- For the COR Social Relations scale, correlations ranged from .34 to .44 (the lowest correlation was with MSCA Verbal and the highest with Motor).
- For the COR Creative Representation scale, correlations ranged from .36 to .52 (the lowest correlation was with MSCA Verbal and Memory and the highest with Perceptual Performance).
- For the COR Music and Movement scale, correlations ranged from .27 to .46 (the lowest correlation was with MSCA Verbal and the highest with Perceptual Performance; this COR scale showed the lowest correlations with MSCA scores).
- For the COR Language and Literacy scale, correlations ranged from .53 to .66 (the lowest correlation was with MSCA Verbal and the highest with Perceptual Performance).
- For the COR Logic and Mathematics scale, correlations ranged from .32 to .46 (the lowest correlation was with MSCA Verbal and the highest with Perceptual Performance; see Schweinhart, McNair, Barnes, & Larner, 1993, p. 452).

Comments

- It is not clear why some of the COR scales (especially Initiative, Social Relations and Creative Representation) would be expected to correlate highly with MSCA scores. The clearest expectations would appear to be for the COR Language and Literacy scale to correlate with the Verbal score of the MSCA; for the COR Logic and Mathematics scale to correlate with MSCA Quantitative score, and perhaps for the COR Music and Movement scale to correlate with the MSCA Motor Skills score. The COR Language and Literacy scale is indeed correlated most highly with the MSCA Verbal score, but the other patterns that seem reasonable to expect did not hold.

- COR does not require that a child be assessed in a context he or she is not familiar with. Rather, progress is charted within a familiar learning environment in the context of ongoing engagement with curricular materials.

- COR does not involve point-in-time assessment, but rather charts the progress over time of the child's engagement in the learning process.

- The reliability of this measure rests on the extent and quality of training and the faithfulness of implementation. It is not clear if initial training is sufficient to achieve adequate reliability or if ongoing training is needed to assure faithfulness of implementation.

- COR is intended to support and inform ongoing instruction. Yet at the same time, the completion of the observations and record keeping take time that could potentially be devoted to instruction.

- As noted by Hoge and Coldarci (1989), reliability is a concern for teacher observation-based measures. The evidence on interrater reliability for this measure is promising, but more extensive study of this issue is warranted (the samples studied here were very small). Further study of interrater reliability with observers of the same general education and experience level (rather than comparing teachers with teaching assistants) would be useful. It would also be important to examine similarity in completion of the ratings across classrooms and age levels in addition to within classrooms.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- In the validation study described above, children's scores on the COR appeared to be weakly correlated with family socioeconomic variables, for example maternal and paternal education were .14 and 28, respectively. Only r-values were given, so it is unknown whether these correlations reached significance.
- It should be noted that this measurement approach was originally created to accompany the High/Scope Curriculum, studied in the High/Scope Preschool Curriculum Comparison Study. However, outcomes on the COR are not reported with the study's results (Schweinhart, Barnes, & Weikart, 1993; Weikart, Bond, & McNeil, 1978).[15].

V. Adaptations of Measure

None found.

[15] See http://www.highscope.org/Research/PerryProject/perrymain.htm for a description of the study.

Early Childhood Measures:
Ongoing Observational

Early Childhood Measures: Ongoing Observational

Creative Curriculum Developmental Continuum for Ages 3-5

I. Background Information

Author/Source

Source: Dodge, D., Colker, L., & Heroman C. (2000). *Connecting content, teaching and learning*. Washington, DC: Teaching Strategies, Inc.

Publisher: Teaching Strategies, Inc.
Box 42243
Washington, DC 20015
Phone: 800-637-3652
Website: www.teachingstrategies.com

Purpose of Measure

As described by instrument publisher:

"The Creative Curriculum Developmental Continuum for Ages 3-5 is an assessment instrument used by teachers to guide them in observing what preschool children can do and how they do it over the course of the year. The Developmental Continuum shows the sequence of development for three-, four-, and five-year-old children on each of the 52 objectives in the Creative Curriculum for Early Childhood. The individual Child Profile shows the developmental indicators for each objective that enable teachers to summarize a child's progress three times a year" (Abbot-Shim, 2001, p.3).

Population Measure Developed With

The Creative Curriculum Developmental Continuum for Ages 3-5 was not developed based on a standardization sample. The measure was developed with, and reliability and validity examined for, the following sample:

- The sample population included 548 children from child care centers (43 percent), Head Start institutions (31 percent), and public preschools (26 percent) located in the northeast, west, south, and southwest United States. Two-thirds of the classrooms were full-day programs.
- The sample was 48.1 percent white, 24.5 percent black, 21.7 percent Hispanic, 4.3 percent Asian/Pacific Islander, .4 percent American Indian/Alaskan Native, and 1 percent other.
- Children ranged in age from 2 years, 8 months to 6 years, 1 month, with a median of 4 years, 4 months. Approximately half were male and half female (52 percent and 48 percent, respectively).
- More than a quarter of the children in the sample spoke a language other than English at home (25.6 percent).

Age Range Intended For

Ages 3 years through 5 years.

Key Constructs of Measure

- The Creative Curriculum Developmental Continuum for Ages 3-5 includes four main constructs: Social/Emotional Development, Physical Development, Cognitive Development, and Language Development.
- Each of the four constructs is broken down into "Curriculum Goals." Each Curriculum Goal consists of individual "Curriculum Objectives." For example, the Social/ Emotional construct has three Curriculum Goals: Sense of Self, Responsibility for Self and Others, and Prosocial Behavior, each with Curriculum Objectives, which are indicators of the development of particular skills. For instance, the Sense of Self Curriculum Goal has four indicators, starting with "Shows ability to adjust to new situations" and ending with "Stands up for rights."
- Each of the Curriculum Objectives is rated by the teacher as "Forerunner, Level I, Level II or Level III." For example, for the Curriculum Objective "Shows ability to adjust to new situations," an example of a Forerunner behavior is "interacts with teachers when family member is around;" an example of a Level I behavior is "says goodbye to family without undue distress;" a Level II behavior is "treats routines and departures as routine parts of the day;" and Level III, "functions independently at school."
- It should be noted that though examples of the various levels are given in the assessment, the fulfillment of these specific examples is not needed, and the end rating is left to the teacher to decide based on his/her observations.

Norming of Measure (Criterion or Norm Referenced)

Criterion referenced.

Comments

- It is not clear whether exceptional children with various mental, physical, or learning disabilities were included in the study sample. The publisher notes (www.teachingstrategies.com) that the inclusion of a Forerunner category may make the measure appropriate for learning delayed students. However it is not clear if this has been examined empirically.

- The instrument's recommended ages for assessment (ages 3 through 5) differ slightly from that of the sample used in the validation study (ages 2 years, 8 months through 6 years, 1 month).

II. Administration of Measure

Who is the Respondent to the Measure?

Teacher.

If Child is Respondent, What is Child Asked to Do?

N/A.

Who Administers Measure/Training Required?

Test Administration:

- Data are collected throughout the school year through multiple methods of assessment such as checklists and anecdotal notes of growth. The teacher observes the child's learning in relation to the goals set by the Creative Curriculum framework. This recorded information is then used to rate children's development on indicators (ratings used are Forerunner, Level I, Level II, or Level III).
- Information about individual children can be rated on the Continuum up to three times a year (fall, winter, and spring), allowing the user to assess change over time.
- The Creative Curriculum system recommends ongoing staff development. To get started in using the curriculum and the assessment tool, a three-day training program is usually needed (www.teachingstrategies.com).

Data Interpretation:

- Interpretation of the Developmental Continuum is fairly straightforward. Those who make the observations should be the ones to do the interpretation. The information can be integrated into daily decisions regarding curriculum and individualization of instruction (www.teachingstrategies.com).

Setting (e.g., one-on-one, group, etc.)

The teacher assesses individual children, but the observation of children may be in a group context.

Time Needed and Cost

Time:

- Ongoing

Cost:

- Curriculum and assessment: $89.95

Comments

- The assessment and curriculum for Creative Curriculum are closely tied. The utility of the assessment apart from the curriculum is not clear.

III. Functioning of Measure

Reliability

Internal Consistency

The scales used in the Developmental Continuum were created through the factor analysis of 52 items, and a four factor solution was found. These factors were then assessed for internal consistency. The coefficient alphas for these factors were .97 for Cognitive Development; .93 for Social Development; .87 for Physical Development; and .91 for Self-Expression. Coefficient alpha for a Total score was .98 (see Abbott-Shim, 2001, p. 9).

Validity

Content Validity

To assess content validity, thirty-nine child development experts reported on whether the items (Curriculum Objectives) of the Developmental Continuum matched the Curriculum Goals, the importance of the items in studying preschool children, and the appropriateness of the Curriculum Objectives as developmental indicators of the Curriculum Goals. There was very little variability in reviewers' responses, most finding Curriculum Objectives important and a good match to Curriculum Goals. When assessing the appropriateness of Curriculum Objectives as developmental indicators of the Curriculum Goals, reviewers generally found them to be appropriate (see Abbott-Shim, 2001, p. 10). We note that this analysis did not address the developmental order of the Curriculum Objectives within the Curriculum goals.

Construct Validity

Construct validity was examined via factor analysis as discussed above.

Comments

- The reliability of this measure rests on the extent and quality of training and the faithfulness of implementation. It is not clear if initial training is sufficient to achieve adequate reliability or if ongoing training is needed to assure faithfulness of implementation.

- Further work on reliability (for example, interrater reliability), and validity (for example, relating scores on this assessment to other assessments, or over time to children's progress in school) would be informative. Reliability is a key issue for observation-based measures (Hoge & Coladarci, 1989), and interrater reliability information for this assessment system is not currently available.

- There are some questions about how the findings from factor analysis were used in defining constructs. Examination of factor loadings indicates that, in some instances, a curriculum goal has been placed with one construct in the assessment when it might better belong with a different construct based on the factor analysis. (See, for example, the curriculum goal of "Reading and Writing," which is placed with the Language Development construct in the Continuum, but loads with the Cognitive Development construct in the validation study.) It would be helpful to have internal reliability estimates for the constructs as they currently stand, as well as an extension of construct validation beyond factor analysis alone. According to the publisher (personal correspondence, 1/7/03) a new validation and reliability study is currently being designed, and results will be posted on the Teaching Strategies website as soon as they are available (www.teachingstrategies.com).

- The content analysis addressed the importance of Curriculum Objectives and their relevance to Curriculum Goals, but not the ordering of Objectives within a developmental sequence. That is, while the current analysis helps to confirm the appropriateness of assigning items to particular Curriculum Goals, it does not reflect on the order in which they are placed.

- It would be helpful to have more information about the functioning of the measure with children who speak a language other than English at home (of particular relevance for Head Start and other early intervention programs), and about differences in the functioning of the measure in full-day and part-day programs.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- One study focused on the curriculum component of Creative Curriculum, rather than the assessment component (Abbott-Shim, 2000). However, as one of the main purposes of the assessment system is to inform curriculum decisions, the study is relevant to the assessment. Using a pre-test/post-test design, the author found significant increases in children's receptive language (PPVT-R; Dunn & Dunn, 1981), WJ-R subtests (Woodcock & Johnson, 1989), and the Adaptive Social Behavior Inventory (Scott, Hogan, & Bauer, 1997) after they had participated in a Creative Curriculum classroom for one year. However, the study did not involve random assignment or a control group.

Comments

- The Creative Curriculum Developmental Continuum does not require that a child be assessed in a context he or she is not familiar with. Rather, progress is charted within a familiar learning environment in the context of ongoing engagement with curricular materials.

- The Creative Curriculum Developmental Continuum does not involve assessment at only a single point in time, but rather charts the progress over time of the child's engagement in the learning process.

- The Creative Curriculum Developmental Continuum is intended to support and inform ongoing instruction. Yet at the same time, the completion of the observations and record keeping take time that could potentially be devoted to instruction.

- Further work regarding the Continuum's usefulness with children who have special needs would be useful.

V. Adaptations of Measure

Spanish Version of Creative Curriculum

A Spanish version of Creative Curriculum is available.

Early Childhood Measures:
Ongoing Observational

Early Childhood Measures: Ongoing Observational

The Galileo System for the Electronic Management of Learning (Galileo)

I. Background Information

Author/Source

Author: Assessment Technology, Inc.

Publisher: Assessment Technology, Inc.
5099 E. Grant Road, Suite 331
Tucson, AZ 85712
Phone: 800-367-4762
Website: www.ati-online.com

Purpose of Measure

As described by instrument publisher:
"In the Galileo System, the major purpose for measuring ability is to promote learning. The developmental perspective introduced by Binet and incorporated into criterion-referenced assessment lends itself well to this purpose. The Galileo System is consistent with Binet's approach. In Galileo, ability is measured in terms of position in an ordered developmental progression. There are two advantages to the Galileo developmental perspective. First, when one knows a child's position in a developmental continuum, it is possible to anticipate the kinds of things that the child will be ready to learn as development progresses. This information is very useful in planning learning experiences to promote growth. The second advantage involves the mathematical models used in Galileo to measure ability. These models make it possible to infer the kinds of things that a child will be capable of doing based on a limited number of observations of what he or she has done. The result is a substantial increase in the amount of information available about children's learning that can be derived from assessment" (The Galileo System Overview, www.ati-online.com, 9/27/02).

Population Measure Developed With

The development of the preschool versions (the earlier MAPS-PL2 and the more recent Galileo) of the measure involved the use of two separate samples of children:

- *1994 Sample*:
  - The MAPS-PL2 Developmental Observation Scales (earlier versions of what is now called Galileo) were developed using a sample of 2,638 children participating in early childhood programs across the country.
  - Children ranged in age from 2 years, 10 months to 5 years, 6 months, and were nearly equally split between male and female.
  - The sample was a fairly close approximation of the 1990 U.S. Census for distribution across the U.S. Thirty-one percent of the children were black, 36 percent white, 30 percent Hispanic, less than 1 percent Asian/Pacific Islander, less than 1 percent American Indian/Alaskan Native, and less than 1 percent other. The 1990 Census had a substantially higher percentage of whites than were included in this sample. Ethnic minorities were over-sampled to better represent

the populations that generally make up Head Start classrooms, and thus do not match 1990 Census data.

- *Fall 2001 Sample*:
  - The Preschool Galileo Assessment Scales were developed using a sample of 3,092 children participating in early childhood programs in the states of Florida, Indiana, Kentucky, Ohio, Oregon, Tennessee, and Texas.
  - Children ranged in age from 3 years, 2 months to 5 years, 10 months.
  - 52 percent of the children were male and 48 percent were female.
  - 43 percent were black, 40 percent were white, and 17 percent were Hispanic.

Age Range Intended For

Birth through age 10, with different scales for different age ranges.

Key Constructs of Measure

- Galileo includes seven scales, each with its own set of constructs:
  - *Infant-Toddler Scales (birth to 2):* Early Cognitive Development, Perceptual-Motor Development, Self-Help, and Social Development.
  - *Preschool Level One Scales (ages 2-4):* Early Math, Language and Literacy, Nature and Science, Perceptual-Motor Development, Self-Help, and Social Development.
  - *Preschool Level Two Scales (ages 3-5):* Approaches to Learning, Creative Arts, Early Math, Fine and Gross Motor Development, Language and Literature, Nature and Science, Physical Health, and Social and Emotional Development.
  - *Kindergarten Level Scales (ages 5-7):* Early Math, Language and Literacy, Nature and Science, and Social Development.
  - *Level One Scales (ages 6-8):* Early Math, Language and Literacy, and Social Development.
  - *Level Two Scales (ages 7-9):* Early Math, Language and Literacy, and Social Development.
  - *Level Three Scales (ages 8-10):* Early Math, Language and Literacy, and Social Development.

(Note: This review focuses on the Preschool Level Two Scales (ages 3-5) in providing examples of content and in discussing reliability and validity. Scale description and reliability and validity information is available for each scale at www.ati-online.com)

The constructs are comprised of "Knowledge Areas." For example, the Early Math construct is comprised of fourteen Knowledge Areas, beginning with the most basic early math skills (e.g., Counting), and moving on to more difficult Knowledge Areas (e.g., Estimation/Comparison, Addition, Subtraction, Fractions).

Each Knowledge Area is measured by a set of indicators, which are arranged in the developmental order in which the children acquire the skills. For example, in the Counting area in the Preschool Level Two Scales, indicators begin with "Uses one-to-one correspondence when counting objects" and end with the more developmentally advanced indicator of "Counts

backward to find how many are left." For this particular Knowledge Area, there are eight indicators (i.e., eight ordered developmental capability levels).

Theory and empirical work were used to develop both the content and sequencing of indicators within Knowledge Areas. For each indicator, the teacher rates the child as exhibiting or not exhibiting the capability.

Norming of Measure (Criterion or Norm Referenced)

Galileo was developed within the framework of criterion-referenced assessments (see Purpose of Measure, above). However, the developer describes the assessment as "path-referenced" (i.e., measuring ability in terms of position in an ordered developmental progression, and the use of a mathematical model to gauge the broader concept of ability; personal correspondence, 6/12/02).

Comments

- The oversampling of ethnic minority groups and the attempt to reflect the Head Start population in the study samples are noteworthy.

- During the development of Galileo, no explicit consideration was given to its usefulness with special populations (e.g., children with learning disabilities, mental retardation, physical or emotional impairment). The potential exists to develop new scales for use with special populations by using a computer program provided with the Galileo system called "Scale Builder," but new scales must be examined for reliability and validity before being implemented.

II. Administration of Measure

Who is the Respondent to the Measure?

Teacher.

If Child is Respondent, What is Child Asked to Do?

N/A.

Who Administers Measure/Training Required?

Test Administration:

- Data are collected throughout the school year, through multiple methods. The teacher observes the child's learning as it relates to Galileo's educational goals. Observational methods can include portfolios, checklists, anecdotal notes, and parental reports. The teacher also completes ratings on the indicators in Galileo's Knowledge Areas.
- The teacher then enters the child's ratings based on what is gleaned from anecdotal notes and checklists into the Galileo computer program. The information can then be analyzed in various ways. For example, Galileo can provide developmental profiles for individual children, inform lesson planning by examining children's progress in relation to educational goals, and aggregate data across children.

- Galileo requires training for ongoing collection of information about individual children, completing child ratings, and using the system's computer program. Training is tailored to fit the needs of a particular school, district, state, etc. Initial training usually takes around two days. Ongoing assistance and technical support are available from the developers and are included within the cost of the system.

Data Interpretation:

- Each time the teacher enters data, Galileo provides a report and curriculum suggestions based on individual and classroom level data.
- All computation is done by the system's computer program. Those with a background in teaching, policy, or education should have no problem interpreting Galileo results.

Setting (e.g., one-on-one, group, etc.)

The teacher makes multiple assessments of individual children, but each child's behavior may be observed in the context of a group.

Time Needed and Cost

Time:

- Ongoing.

Cost:

- Galileo Online (online version): $300 per class, per year, no start-up fee, includes program updates and tech support.
- Galileo G2 (stand alone version): $370 per class, per year, includes online aggregation, tech support, and updates.

Comments

- A computer is needed to use Galileo. Teachers and early childhood programs differ in their access to and comfort with using computers.

- The program itself is designed to be "user friendly." For example, analyses are done by the program for the user. However, understanding the statistical methods underlying the system requires a background in psychometrics.

III. Functioning of Measure

Reliability Information from Manual

Internal Consistency

The Manual reports on internal consistency for each construct using marginal reliability coefficients. "The marginal reliability coefficient combines measurement error estimated at different points on the ability continuum into an overall reliability coefficient, which corresponds closely to conventional estimates of reliability" (www.ati-online.com, 9/27/02). For the Preschool Level Two Scales (ages 3-5), marginal reliability coefficients were as follows: Approaches to Learning was .94; Creative Arts was .96; Early Math was .95; Fine and Gross

Motor was .92; Language and Literacy was .97; Nature and Science was .97; Physical Health was .96; and Social and Emotional was .97 (see www.ati-online.com).

Interrater Reliability
Observations were conducted on three randomly selected children from each of 318 classrooms in three Ohio Head Start programs. An onsite coordinator was designated in each program to supervise data collection for a primary observer (the lead teacher) and secondary observer (assistant teacher) for each classroom involved in the study.

After a substantial training period, the teachers and assistant teachers each gave the three observed children in their classrooms scores on both the Early Math and the Language and Literacy constructs. Each observer averaged the Early Math scores of the three children that he/she assessed, and the same was done for Language and Literacy scores. The correlation between the two observers' (i.e. teacher and assistant teacher) average Early Math scores and average Language and Literacy scores were assessed to yield a measure of observer agreement at the classroom level. The class level was taken as the unit of analysis because the data provided to the state are aggregated, and the class is the smallest unit of aggregation. Agreement was high for the two scales assessed: correlations averaged .83 for Early Math, and .88 for Language and Literacy (see www.ati-online.com).

Validity Information from Manual
Content Validity
Galileo bases its content validity on regularly updated literature reviews. The literature provides the foundation for the identification of Knowledge Areas (e.g., Counting), for the developmental sequencing of indicators within Knowledge Areas, and for the ordering of Knowledge Areas themselves in a developmental sequence. The formulation based on the review of the literature is then examined empirically.

Construct Validity
The indicators within each Knowledge Area were examined for developmental order and cohesiveness, using Item Response Theory, an approach based in Latent Trait statistical methods (see the Galileo Online Manual [www.ati-online.com] for a full overview, justification, and history of these methods).

For each Knowledge Area, a "Discrimination Value" was calculated to estimate the degree to which the capabilities rated in the indicators are related to the underlying abilities being measured. All Knowledge Areas had acceptable discrimination values

Similarly, a "Difficulty Value" was computed to determine the position of individual indicators within the Knowledge Areas and assess whether the indicators are in proper developmental order. Developmental order, in this case, is established using IRT methods to model the order in which children accomplished the indicators in the sample. Difficulty values are a way to quantify whether the developmental order established by the IRT model are valid. The difficulty value for each scale was reported to show acceptable developmental progress (www.ati-online.com, 9/27/02).

Reliability/Validity Information from Other Studies
None found.

Comments

- The reliability of this measure rests on the extent and quality of training and the faithfulness of implementation. It is not clear if initial training alone is sufficient to achieve adequate reliability or if ongoing training is needed to assure faithfulness of implementation.

- The measure of interrater reliability summarizes scores for individual children within a classroom. It would be useful to have a measure of interrater reliability based on scores for individual children.

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

None found.

Comments

- Galileo does not require that a child be assessed in a context he or she is not familiar with. Rather, progress is charted within a familiar learning environment in the context of ongoing engagement with curricular materials.

- Galileo does not involve point-in-time assessment, but rather charts the progress over time of the child's engagement in the learning process.

- Galileo is intended to support and inform ongoing instruction. Yet at the same time, the completion of the observations and record keeping take time that could potentially be devoted to instruction.

The Online Manual provides an extensive discussion of Galileo's theoretical and empirical approach.

The developmental sequence of indicators and Knowledge Areas was based on a literature review as well as empirical examination. This approach is noteworthy because an empirical examination of sequencing is lacking in some other ongoing observational measures.

The use of this assessment system requires access to a computer and training in use of the computer program. While training and use do not seem overly difficult, this may be a barrier in some settings.

V. Adaptations of Measure

"Scale Builder"

Galileo "Scale Builder" allows the system to be translated into other languages.

Early Childhood Measures:
Ongoing Observational

Early Childhood Measures: Ongoing Observational

The Work Sampling System (WSS)

I. Background Information

Author/Source

Source: Meisels, S., Jablon, J., Dichtelmiller, M., Dorfman, A., & Marsden, D. (1998). *The Work Sampling System*. Ann Arbor, MI: Pearson Early Learning.

Meisels, S., Bickel, D., Nicholson, J., Xue, J., Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal, 38*(1), 74-95.

Publisher: Pearson Early Learning
P.O. Box 2500
135 South Mt. Zion Road
Lebanon, IN 46052
Phone: 800-321-3106
Website: www.pearsonearlylearning.com

Purpose of Measure

As described by instrument publisher:
"The Work Sampling System is a validated, research-based observational assessment designed to enhance instruction and improve learning for preschool to grade 6. The Work Sampling System 4th Edition reflects the recent changes in standards and assessment. It focuses clearly on high standards of learning and instructionally meaningful, developmentally appropriate teaching. Work Sampling provides insight into how an individual child learns and targets the following areas: Personal and Social Development, Language and Literacy, Mathematical Thinking, Scientific Thinking, Social Studies, The Arts, and Physical Development and Health" (from Website; see www.pearsonearlylearning.com).

Population Measure Developed With

- This measure was not developed using a standardization sample. Rather, the measure charts growth and development over time in terms of specific criteria.
- Reliability and validity for the most recent version of WSS were assessed with the following sample of children:
    - The sample was taken from five public schools located in Pittsburgh where WSS had been implemented for three years and included 17 teachers who had had at least two years of experience using WSS.
    - The sample consisted of 345 children in four cohorts: kindergarten (N = 75), first grade (N = 85), second grade (N = 91), and third grade (N = 94).
    - Race/ethnicity of the children in the sample included black, white, Hispanic, Asian/Pacific Islander, and other. The largest representation in each cohort was black. Composition varied somewhat by cohort, with the third grade cohort

having relatively more black children and fewer white children than the other cohorts, and the kindergarten cohort having a greater representation of Asian children (see Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001, p. 78).
    - The sample was 48.7 percent male; 7.8 percent of the children were classified as having special needs; and 79.4 percent received reduced cost or free lunches (a proxy for household income).

Age Range Intended For
Ages 3 years through Grade 6.

Key Constructs of Measure

- The WSS focuses on seven constructs ("domains"):
    - *Personal and Social Development*: child's feelings about self and interactions with peers and adults.
    - *Language and Literacy*: acquisition of language and reading skills.
    - *Mathematical Thinking*: patterns, relationships, the search for multiple solutions to problems. Both the aspects of *concepts and procedures* and *knowing and doing* are addressed.
    - *Scientific Thinking*: how children investigate through observing, recording, describing, questioning, forming explanations, and drawing conclusions.
    - *Social Studies*: ideas of human interdependence and the relationships between people and the environment.
    - *The Arts*: how children engage in dance, drama, music and art, both actively and receptively.
    - *Physical Development*: addresses fine and gross motor development, control, balance and coordination.

Each of these domains includes "Functional Components." For instance, the Language and Literacy construct is broken down into the following Functional Components: Listening; Speaking; Literature and Reading; Writing; and Spelling. Each of the Functional Components is defined by a series of performance indicators that present "the skills, behaviors, attitudes and accomplishments" of the child (Dichtelmiller, Jablon, Meisels, Marsden, & Dorfman, 1998, p. 11).

Norming of Measure (Criterion or Norm Referenced)
Criterion referenced.

Comments

- The sample used for examining reliability and validity was not representative of the U.S. population. However, versions of WSS are being used statewide in a number of states for children in specific grades (e.g., South Carolina, Maryland), and data on the use of WSS with wider demographic ranges is likely forthcoming.

- The development of the measure with a sample of low-income children, many from minority racial/ethnic groups, suggests that this measure is appropriate for use in Head Start and other early intervention programs.

- Information on reliability and validity of the most recent version of the measure was not found for the full age range that the measure is designed for (information was found for kindergarten through third grade). It is possible that additional data are forthcoming or that we have not located the data.

- As noted by Meisels and colleagues (Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001), WSS was initiated in the Pittsburgh School District as part of a district-wide restructuring to improve student outcomes. At the same time that WSS was implemented, so were new reading and social studies programs in the elementary grades and a new math program in the third grade. This context needs to be taken into account when considering results. WSS as a tool to improve instruction cannot be isolated from the influence of these other changes.

II. Administration of Measure

Who is the Respondent to the Measure?
Teacher.

If Child is Respondent, What is Child Asked to Do?
N/A.

Who Administers Measure/ Training Required?
Administration

- Data are collected throughout the school year, through portfolios, developmental guidelines, and checklists and then compiled in summary reports.
- Portfolios in WSS are used to track a child's efforts, achievements, and progress, and are designed to do this in two ways: by collecting student work that reflects "Core Items" (items that show growth over time within a given construct), as well as "Individualized Items" (items that reflect unique aspects of the development of the child that cross over multiple constructs).
- Developmental checklists are provided for each construct. These include a brief description of the developmental expectations for the Functional Components of the construct being addressed, and a few examples of how the one-sentence indicator might be met. The specific indicator is then rated in a trichotomous fashion: Not Yet, In Progress, or Proficient. For instance, within the construct of Language and Literacy, one Functional Component is "Listening." Within the Listening component for first grade children, a brief description of what listening skills could be expected of a 6-year-old child is provided. Following this description, examples are given of how a child might display the behavior, such as "Child asks a relevant question of a friend regarding the story the friend conveyed." The teacher is then required to rate the level of the child's

skills on a particular indicator, such as "Listens for meaning in discussions and conversations" as Not yet, In Progress, or Proficient.

- A Summary Report is to be prepared three times a year (replacing conventional report cards). Each Functional Component is rated for Performance (Developing as Expected, or Needs Development) for both checklists and portfolios, as well as for Progress (As Expected, or Other Than Expected). Teachers can also add comments to the ratings.

Training

- Training to use the WSS is available through on-site consultations or national workshops lasting from one to three days.

Data Interpretation

- The teachers who maintain the records should also interpret the results and use them on an ongoing basis to inform instruction.

Setting (e.g., one-on-one, group, etc.)

The teacher assesses the progress of individual children, but the children can be observed in groups as well as individually in the classroom.

Time Needed and Cost

Time

- Ongoing.

Cost

- Starts at $75 (for the basic Teacher Reference Pack) and increases in price depending on materials needed. Sections can be purchased separately.

Comments

- While it only takes one to three days to learn how to implement the WSS, what is gleaned from the training and how it is applied might vary depending on teacher experience and training.

III. Functioning of Measure

Reliability

Internal Consistency

Internal consistency was reported for an earlier version of WSS, based on use with a sample of kindergarten children from ethnically and economically varying communities in Michigan. Coefficient alphas ranged from .87 to .94 on checklist scales for the final of three waves of testing that were done: Art & Fine Motor = .87, Movement & Fine Motor = .91, Concept & Number = .91, Language & Literacy = .94, and Personal/Emotional Development = .93 (see Meisels, Liaw, Dorfman, & Nelson, 1995, p. 287). We did not find results on internal consistency in the published report regarding the most recent edition of WSS (Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001).

Interrater Reliability
Similarly, no interrater reliability was reported for the most recent edition of the WSS, but was reported for the earlier version. In the case of the earlier WSS, two raters were asked to complete the WSS summary report for 24 familiar children and 25 unfamiliar children based on the children's portfolios and checklists. Correlations for ratings by the two raters were high (r = .88). Correlations between the ratings of the two raters and the children's teachers were lower but still high (.73 and .68; see Meisels, Liaw, Dorfman, & Nelson, 1995, p. 291).

Validity
Concurrent Validity
Data on validity were collected for the current version of WSS (Meisels et al., 2001) for a sample of 345 children from 17 classrooms in schools in Pittsburgh. The children were broken into four cohorts—kindergarten, first, second, and third grade (for further description, see "Population Measure Developed With," above). In addition to the checklist and Summary Report ratings from WSS, each student in the sample was assessed with the Woodcock Johnson-Revised (WJ-R; Woodcock & Johnson, 1989) battery. Correlations between WJ-R standard scores for specific subscales and the WSS Language and Literacy checklist, the WSS Mathematical Thinking checklist, and Summary Report ratings were assessed. Correlations between the most relevant WJ-R subscales and WSS checklist and Summary Report ratings at two time points (Fall and Spring) ranged from .36 to .75, with the majority of coefficients falling between .50 and .75. Correlations tended to increase with age, but varied depending upon WJ-R scale. Relationships between WJ-R scales and WSS scores were consistently stronger in the spring assessment of the WSS for every age group other than third grade (see Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001, p. 82-83).

Unique contributions of the WSS checklist scores in predicting WJ-R standard scores, beyond age, SES, ethnicity, and initial performance on the WJ-R, were assessed through multiple regression analysis. In kindergarten and first grade (though not in second and third grade), WSS checklist scores were significantly related to WJ-R math scores, with the other variables (including initial WJ-R score) taken into account. Similarly, for children in kindergarten through second grade (though not third grade), WSS checklist scores were related to WJ-R language and literacy scores even after the other variables were taken into account. It is noted that in the later grades, when standardized scores usually become more stable, initial WJ-R scores accounted for almost half of the variability in the later WJ-R scores.

WSS Summary Report scores were significantly related to WJ-R language and literacy scores for kindergarten, first, and second grade. Similar patterns were found for the WSS Math checklists and Summary Reports with respect to scores on WJ-R Mathematical Thinking scores.

Using the same data as noted above for studying the concurrent validity of WSS ratings, cut-offs were created to identify "at risk" and "not at risk" scores on both the WJ-R and on WSS Broad Reading and Broad Math. A student shown to be at-risk in either reading or math on WSS "has a much higher probability of being ranked lower on the WJ-R than a randomly chosen student who is performing at or above average" (Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001. p. 88).

Comments

- The regression results, taking into account demographic variables and earlier scores on the WJ-R, are further examples of the acceptable concurrent validity of the WSS for grades other than third grade. As the authors note, the lack of significant associations for third grade children may reflect the greater stability of standardized assessment scores for older children. Further work on validity especially at the older end of the age range is warranted.

- We were not able to locate any information on the reliability of the current version of WSS. Though a study of an earlier version of the measure did report such information, the scales and populations were different than those for the current version. This is an important issue given the fact that reliability is a concern for teacher observation-based measures (Hoge & Coladarci, 1989).

- The examination of interrater reliability using the earlier version of WSS leaves questions open. The two raters based their scores on portfolios and checklists collected by a teacher throughout the year (i.e., existing information). The possibility remains that teachers differing in experience, training and/or beliefs might agree on how to rate existing information, but differ in terms of what they would deem relevant to collect in terms of portfolios, or how they would complete checklists. In addition, it is noteworthy that "rater-rater" agreement was stronger than "rater-teacher" agreement.

- We also found no information regarding content validity, most notably a rationale for how WSS developers identified behavior for the Functional Components for each age, save the indication that they were based on " learner-centered expectations that were derived from national and state curriculum standards" (Meisels, et al., 2001, p. 78). It would be helpful to have an articulated justification for the choice of Functional Components and behavioral indicators of development.

- Reliability and validation information is currently available only for children between kindergarten and third grade, although the publisher reports that the WSS is appropriate for children from age 3 to grade 6. According to one of the WSS authors, information for the further age ranges is forthcoming (personal communication, 1/16/03).

IV. Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- Though the WSS curriculum/assessment was used as a predictor, as opposed to an outcome in a study done by Meisels, Atkin-Burnett, Xue, Nicholson, Bickel, & Son (in press), the findings remain relevant given that ongoing WSS assessment is an integral part of the curriculum. In a natural experiment (i.e., not randomly assigned to treatment or control), a group of children in a low-income urban setting received three years of the WSS curriculum and assessment. This group was then compared against two non-WSS groups, one matched for demographics and the other representing the remainder of the

students within the city (PPS). Each group was given the Iowa Test of Basic Skills (ITBS) in the third and forth grade, and mean change scores for each of these groups were used as the dependent variable. The authors found that the children who received the WSS showed greater change scores from third to forth grade on ITBS rated reading than both the matched sample and PPS groups. A similar relationship was found for ITBS math scores, with WSS change scores marginally larger than the PPS group, and significantly larger than the matched sample. It should be noted that while the WSS may show a relationship to change scores using the ITBS, this was not a controlled study. Various other curricula changes were made simultaneous to the adoption of WSS, adoption of WSS, in itself, was a voluntary choice made by the teacher, and after independent review of the classroom, only the best classrooms rated as having high WSS implementation standards were included in the analysis. That is, selection effects might explain the current results.

Comments

- WSS does not require that a child be assessed in a context he or she is not familiar with. Rather, progress is charted within a familiar learning environment and in the context of ongoing engagement with curricular materials.

- The reliability of this measure rests on the extent and quality of training and the faithfulness of implementation. It is not clear if initial training is sufficient to achieve adequate reliability or if ongoing training is needed to assure faithfulness of implementation.

- WSS does not involve point-in-time assessment, but rather charts the progress over time of the child's engagement in the learning process.

- WSS is intended to support and inform ongoing instruction. Yet at the same time, the completion of the observations and record keeping take time that could potentially be devoted to instruction.

V. Adaptations of Measure

There are Spanish language versions of some of the WSS materials.

Ongoing Observational Measures References

Abbott-Shim, M. (2001). *Technical report: Validity and reliability of The Creative Curriculum for Early Childhood and Developmental Continuum for Ages 3-5*. Atlanta, GA: Quality Assist, Inc.

Abbott-Shim, M. (2000*). Sure Start Effectiveness Study: Final report*. Atlanta, GA: Quality Assist, Inc

Assessment Technology, Inc. (2002) *Galileo online technical manual.* Retrieved from http://www.assessmenttech.com/pages/research/galileotechmanual_files/contents.html.

Dodge, D., Colker, L. & Heroman C. (2000). *Connecting content, teaching and learning*. Washington, DC: Teaching Strategies, Inc.

Dichtemiller, M., Jablon, J., Meisels, S., Marsden, D., Dorfman, A. (1998). *Using work sampling guidelines and checklists: An observational assessment*. Ann Arbor, MI: Rebus, Inc.

Dunn, L., & Dunn, L. (1981). *Peabody Picture Vocabulary Test – Revised*. Circle Pines, MN: American Guidance Service.

Epstein, A.S. (1992). *Training for quality: Improving early childhood programs through systematic in-service training.* (Final report of the High/Scope Training of Trainers Evaluation). Ypsilanti, MI: High/Scope Educational Research Foundation.

High/ Scope Educational Research Foundation (2002): http://www.highscope.org/

Hoge, R.D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research, 59*(3), 297- 313.

McCarthy, D. (1972). *McCarthy Scales of Children's Abilities*. San Antonio, TX: The Psychological Corporation.

Meisels, S. Bickel, D., Nicholson, J., Xue, J., Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal, 38(1)*, 74-95.

Meisels, S., Jablon, J., Dichtelmiller, M., Dorfman, A., & Marsden,D. (1998) *The Work Sampling System.* Ann Arbor: MI: Pearson Early Learning.

Meisels, S., Liaw, F. , Dorfman, A., Nelson, R. (1995). The Work Sampling System: Reliability and validity of a performance assessment for young children. *Early Childhood Research Quarterly, 10,* 277-296.

Schweinhart, L., Barnes, H., Weikart, D. (1993). Significant benefits: The High/Scope Perry Preschool Study. *Monographs of the High/Scope Educational Research Foundation 10.* Ypsilanti, MI: High/Scope Educational Research Foundation.

Schweinhart, L., McNair, S., Barnes, H., & Larner, M. (1993). Observing young children in action to assess their development: The High/Scope Child Observation Record Study. *Educational and Psychological Measurement, 53,* 445-454.

Scott, K.G., Hogan, A., & Bauer, C. (1997). Social Competence: The Adaptive Social Behavior Inventory (ASBI). In R.T. Gross, D. Spiker, & C.W. Haynes (Eds.), *Helping Low Birth Weight, Premature Babies: The Infant Health and Development Program*. Stanford, CA: Stanford University Press

Woodcock, R.W. & Johnson, M.B. (1989). *Woodcock-Johnson Psycho-Educational Battery – Revised*. Itasca, IL: Riverside Publishing.

Early Head Start[16]

The following is a list of measures used in the pre-kindergarten follow-up of the Early Head Start Research and Evaluation project. These include both cross-site measures (indicated with an asterisk), as well as a partial list of those used in site-specific research. The cross-site measures overlap substantially with FACES measures. Not all of the measures listed below are included in the compendium of measurement descriptions provided.

Social Emotional

- Child Behavior Checklist (CBCL) Aggression subscale *
- Howes Peer Play scale *
- The Moral Dilemma Situation (Buchsbaum and Emde, 1990)
- MacArthur Story Stem Battery
- What I think about school instrument (Ramey, 1988)
- Attachment Q-set
- Penn Interactive Peer Play Scale (PIPPS)

Cognitive

- Leiter R - sustained attention task and observation ratings *
- Woodcock-Johnson Applied Problems *
- Kaufman Brief Intelligence Test
- Theory of mind tasks—false identity task and false location task

Mastery

- Dimensions of Mastery Questionnaire (Morgan et al, 1990)

Language

- Peabody Picture Vocabulary Test - Third Edition *
- Woodcock-Johnson Letter-Word Identification *
- Modified Story and Print Concepts *
- Observations of parent/child book reading scored using the Child Language Data Exchange System
- Rhyme and Deletion tasks of the Early Phonemic Awareness Profile
- TOLD Phonemic Analysis Subscale
- Minnesota Literacy Indicator

Parenting

[16] *The information contained in the document was gathered by the Child Outcomes Research and Evaluation/Office of Planning, Research, and Evaluation (CORE/OPRE) in ACF

- Child Abuse Potential Inventory
- Parenting Stress Inventory
- Stony Brook Family Reading Survey

Parent Mental Health/Family Functioning
- CES-D *
- Impacts of Events Scale
- Family Crisis Oriented Personal Scale (F-COPES; McCubbin, 1987)
- Brief Symptom Index
- Dyadic Adjustment Scale

Quality of the Home Environment
- Home Observation for Measurement of the Environment (HOME) *

Quality of the Child Care Setting
- ECERS/FDCRS
- Arnett

Note: A complete list of cross-site measures from the birth to three study can be found at: http://www.acf.dhhs.gov/programs/core/ongoing_research/ehs/ehs_instruments.html

Early Childhood Workshop

Early Childhood Workshop

Child-Parent Rating Scales for the Puzzle Challenge Task

Prepared by Christy Brady-Smith
National Center for Children and Families, Teachers College, Columbia University

I. Background Information

Author/Source

Author: Christy Brady-Smith, Rebecca Ryan, Lisa J. Berlin, Jeanne Brooks-Gunn, & Allison Fuligni (2001)
Publisher: unpublished scales, National Center for Children and Families, Teachers College, Columbia University

The scales were adapted from the *Manual for Coding the Puzzle Task* from the Newark Observational Study of the Teenage Parent Demonstration (Brooks-Gunn, Liaw, Michael, & Zamsky, 1992)

The Puzzle Challenge Task was based on the work of Matas, Sroufe, and colleagues (Matas, Arend, & Sroufe, 1978; Sroufe, Egeland, & Kreutzer, 1990).

Purpose of Measure

- Assess child and parent behaviors and child-parent interaction during a task that is challenging for the child
- Overcome possible biases of self-report parenting measures and lab settings by videotaping interaction in the home

Population Measure Developed With

- Participants were low-income White (41%), Black (35%), and Latina (24%) dyads participating in the Early Head Start Research and Evaluation Project

Age Range Intended For

- 36-month-old child and his/her parent

Key Constructs of Measure

- *Child constructs*:
  - Engagement of parent (extent to which child initiates and/or maintains interaction with parent)
  - Persistence (degree to which child is goal-oriented, focused and motivated to complete the puzzles)

    - Frustration with task (degree to which child shows anger or frustration with the puzzle task)
- *Parent constructs*:
    - Supportive presence (the degree to which the parent provides emotional, physical, and affective support to the child during the task)
    - Quality of assistance (the quality of instrumental support and assistance the provided to the child)
    - Intrusiveness (over-involvement, over-control)
    - Detachment (under-involvement and lack of awareness, attention, engagement)
- Constructs assessed on a seven-point scale: "1" indicating a very low incidence of the behavior and "7" indicating a very high incidence of the behavior
- Contact the National Center for Children and Families (nccf@tc.columbia.edu) for additional information on the coding scales

Norming of Measure (Criterion or Norm Referenced)

- Training tapes for the videotape coders included examples of supportive and unsupportive or intrusive parenting behaviors for all three racial/ethnic groups.
- Coders' reliability tapes were randomly assigned and included White, Black, and Latina dyads
- Preliminary analyses examined inter-scale correlations, possible underlying factors, and internal consistency for the full sample and by race/ethnicity, and scales appeared to be operating similarly for all groups. Future analyses with examine these issues further.

Concerns, Comments, and Recommendations:

II. Administration of Measure

Who is the Respondent to the Measure?

- The child and parent

If Child is Respondent, What is Child Asked to Do?

- The child is asked to solve up to three puzzles of increasing difficulty in 6 to 7 minutes. The parent is instructed to let the child work on the puzzle independently first and then give the child any help he or she may need. The dyad has up to four minutes to work on each puzzle. If the child does not solve the first puzzle in four minutes, the interviewer/assessor asks the child to try the second puzzle.
- The first puzzle is relatively easy (9 pieces in easy-to-fit positions), the second puzzle is more difficult (10 pieces), and the third puzzle is quite challenging (20 pieces).

Who Administers Measure/ Training Required?

Test Administration

- The protocol was administered by trained interviewer/assessors (I/As)
- Training sessions for I/As were held at Mathematica Policy Research, Inc. (MPR) and conducted by MPR staff
- I/As also were responsible for videotaping the dyad and keeping distractions to a minimum by asking other family members to leave the area

Data Coding

- At Columbia University, a team of six graduate students was trained to code the videotaped vignettes.
- Training included weekly meetings, discussions of the scales, and viewing of the training tapes that contained exemplars of high, medium and low scoring interactions for each scale.
- Coders reached 85% agreement (exact or within one point) or higher with a "gold standard" before coding unique interactions.
- A randomly selected 15% to 20% of each coder's weekly tape assignments were used to ensure ongoing reliability.
- Coders were ethnically heterogeneous
- Interactions conducted in Spanish were rated by a fluent Spanish-speaking coder
- Coders were unaware of participants' treatment group status

Setting (e.g. one on one, group, etc):

- Child-parent interactions were videotaped in the home

Time Needed and Cost

- 8 to 9 minutes
- Estimated cost of graduate student training and videotaped coding is $95 per videotape

Concerns, Comments, and Recommendations:

- Most children did not get to the third puzzle in the allotted time. The second puzzle (black and white panda bears) seemed to be very challenging to the children.

III. Functioning of Measure

Reliability

Coder Reliability – Percent agreement (exact or within one point) averaged 93% across all 36-month puzzle task coders, with a range of 88% to 100%.

- A total of 194 tapes (12 percent of the 1,639 codable tapes) served as reliability tapes

Internal Reliability – Not assessed as there were no composite variables. The correlation among child engagement and frustration with the task was not significant (r = -.05); correlations among the other child scales were moderate to high (statistically significant $|r|$'s = -.21 to .41). The correlations among the four parenting scales were moderate to high and statistically significant ($|r|$'s = -.27 to .59), with the exception of the correlation between intrusiveness and detachment, which was small but significant (r = .16).

Validity

- Several papers have been proposed by the Early Head Start Consortium Parenting Workgroup to explore the validity of this measure
- The observational measures will be compared to widely-used assessments tapping similar parenting (e.g., HOME) and child constructs (e.g., Bayley, MacArthur, CBCL)

Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

Concerns, Comments & Recommendations

- The puzzle challenge task has not been used in a wide variety of studies and has not been subject to vigorous psychometric testing
- Researchers in the Early Head Start Consortium have proposed to explore the validity of this assessment
- One promising element of the task is that it elicits a wider range of negative child behaviors compared to the free play (Three Bag) task

IV. Adaptations of Measure

Adaptation:

- N/A

Description of Adaptation

Psychometrics of Adaptation

Study Using Adaptation

Early Childhood Workshop

Child-Parent Interaction Rating Scales for the Three-Bag Assessment

Prepared by Rebecca Fauth and Christy Brady-Smith
National Center for Children and Families, Teachers College, Columbia University

I. Background Information

Author/Source

14-month coding scales:
Author: Anne Ware, Christy Brady-Smith, Claudia O'Brien, & Lisa Berlin (1998)
Publisher: unpublished scales, National Center for Children and FamiliesNational National Center for Children and Families, Teachers College, Columbia University

24-month coding scales:
Author: Christy Brady-Smith, Claudia O'Brien, Lisa Berlin, & Anne Ware (1999)
Publisher: unpublished scales, National Center for Children and Families, Teachers College, Columbia University

36-month coding scales:
Author: Christy Brady-Smith, Claudia O'Brien, Lisa Berlin, Anne Ware, & Rebecca C. Fauth (2000)
Publisher: unpublished scales, National Center for Children and Families, Teachers College, Columbia University

NOTE: The above scales were modified from the *Mother-Child Interaction Rating Scales for the Three-Box Procedure* used in the NICHD Study of Early Child Care (NICHD Early Child Care Research Network, 1997; 1999; Owen, 1992; Owen, Norris, Houssan, Wetzel, Mason, & Ohba, 1993) and the *Manual for Coding Freeplay--Parenting Styles* used in the Newark Observational Study of the Teenage Parent Demonstration (TPD; Brooks-Gunn, Liaw, Michael, & Zamsky, 1992; Spiker, Ferguson, & Brooks-Gunn, 1993).

Purpose of Measure

- Assess child and parent behaviors and child-parent interactions during a semi-structured free play task in a home setting
- Overcome possible biases of self-report parenting measures by using videotaped interactions

Population Measure Developed With

- Participants were low-income White (41%), Black (35%), and Latina (24%) dyads participating in the Early Head Start Research and Evaluation Project

Age Range Intended For

- 14-, 24-, and 36-month child and his/her parent
- Currently adapting the assessment to use for pre-K and first grade children

Key Constructs of Measure

- *Child Constructs*:
- Engagement of parent (extent to which child shows, initiates, and/or maintains interaction with parent
- Sustained attention (degree to which child is involved with toys presented in three bags)
- Negativity toward parent (degree to which child shows anger, hostility, or dislike toward parent)
- *Parent Constructs*:
- Sensitivity* (degree to which parent observes and responds to child's cues)
- Positive regard* (expression of love, respect, and/or admiration for child)
- Stimulation of cognitive development* (quality and quantity of parent's effortful teaching to enhance child's development)
- Detachment (lack of awareness, attention, and engagement with child)
- Intrusiveness (degree to which parent exerts control over child)
- Negative regard (expression of discontent with, anger toward, disapproval of, and/or rejection of child)
- *Sensitivity, positive regard, and stimulation of cognitive development were collapsed into a single scale, Supportiveness, by computing the mean of the three items which were highly intercorrelated (*r*'s = .50 to .71 at all time points)
- Constructs assessed on a seven-point scale, "1" indicating a very low incidence of the behavior and "7" indicating a very high incidence of the behavior
- Contact the National Center for Children and Families (nccf@tc.columbia.edu) for additional information on the coding scales

Norming of Measure (Criterion or Norm Referenced)

- Training tapes for the videotape coders included examples of sensitive/supportive and insensitive parenting behaviors for all three racial/ethnic groups
- Coders' reliability tapes were randomly assigned and included White, Black, and Latina dyads
- Preliminary analyses examined inter-scale correlations, possible underlying factors, and internal consistency for the full sample and by race/ethnicity, and scales appeared to be operating similarly for all groups

Concerns, Comments, and Recommendations:

II. Administration of Measure

Who is the Respondent to the Measure?

- The child and parent

If Child is Respondent, What is Child Asked to Do?

- Child and parent are presented with three bags of toys (labeled #1, #2, and #3, respectively) and are asked to spend 10 minutes with the toys in the three bags beginning with the first bag and ending with the third bag
- Parents may play with the child if they choose
- Contents of the three bags varied according to the age of the child:
    - 14-month children:
        - Bag #1: *Good Dog Carl* book
        - Bag #2: stove, pots, pans, and utensils set
        - Bag #3: Noah's Ark and animals
    - 24-month children:
        - Bag #1: *The Very Hungry Caterpillar* book
        - Bag #2: stove, pots, pans, and utensils set
        - Bag #3: Noah's Ark and animals
    - 36-month children:
        - Bag #1: *The Very Hungry Caterpillar* book
        - Bag #2: groceries, shopping basket, and cash register
        - Bag #3: Duplo blocks

Who Administers Measure/ Training Required?

Test Administration

- The protocol was administered by trained interviewer/assessors (I/As)
- Training sessions for I/As were held at Mathematica Policy Research, Inc. (MPR) and conducted by MPR staff
- I/As also were responsible for videotaping the dyad and keeping distractions to a minimum by asking other family members to leave the area

Data Coding

- At Columbia University, small teams of 5 to 6 graduate students were trained to view and code each videotaped vignette
- Training included weekly meetings, discussions of the scales, and viewing of the training tapes that contained exemplars of high, medium and low scoring interactions for each scale
- Coders reached 85% agreement (exact or within one point) or higher with a "gold standard" before coding unique interactions
- A randomly selected 15% to 20% of each coder's weekly tape assignments were used to ensure ongoing reliability

- Coders were ethnically heterogeneous
- Interactions conducted in Spanish were rated by a fluent Spanish-speaking coder
- Coders were unaware of participants' treatment group status

Setting (e.g. one on one, group, etc):
- Child-parent interactions were videotaped in the home

Time Needed and Cost
- 12 to 13 minutes
- Estimated cost of graduate student training and videotape coding is $95.00 per videotape

Concerns, Comments, and Recommendations:
- Anecdotal evidence from coders and I/As indicates that non-English-speaking parents may have viewed the task differently than those who were more acculturated. Methodology papers designed to address this issue have been proposed by members of the Early Head Start Consortium Methods Workgroup.

III. Functioning of Measure

Reliability

Coder Reliability -- Percent agreement (exact or within one point) averaged 90% at 14 months, with range of 83% to 97%; averaged 93% at 24 months, with a range of 84% to 100%; and averaged 94% at 36-months, with a range of 86% to 100%
- A total of 215 tapes (11% of 1,976 codable tapes) at 14-months, 151 tapes (9% of 1,782 codable tapes) at 24-months, and 174 tapes (11% of the 1,660 codable tapes) at 36-months served as reliability tapes

Internal Reliability -- For the composite supportiveness scale, alpha coefficients ranged from .82 to .83 over the three waves. The correlations among the four parenting scales (supportiveness, intrusiveness, negative regard, and detachment) were small to moderate and statistically significant ($|r|$'s = .11 to .40 at 24 months and .12 to .36 at 36 months), with the exception of supportiveness and detachment ($|r|$'s = .56 and .45, respectively) and intrusiveness and negative regard ($|r|$'s = .52 and .47, respectively).

Validity
- Several papers have been proposed by the EHS Consortium Parenting and Methods Workgroups to explore the validity of this measure
- The parent and child observational measures will be compared to widely-used assessments that tap similar parenting (e.g., HOME) and child constructs (e.g., Bayley, MacArthur CDI, CBCL)

Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- NICHD Study of Early Child Care (NICHD Early Child Care Research Network, 1997; 1999) examined several correlates of supportive parenting, including maternal depression and child outcomes. Parenting also appeared to buffer the effect of low-quality child care on child outcomes.
- Newark Observation Study of the Teenage Parent Demonstration (TPD). TPD is an intervention. The "enhanced services group" of the Teenage Parent Demonstration (TPD) required young mothers to participate in work-related activities, offered moderate support services, and imposed sanctions for non-compliance. Compared to mothers who were not subject to these requirements (due to random assignment), mothers in the enhanced-services group were less responsive with their children (Kisker, Rangarajan, & Boller, 1998).

Concerns, Comments & Recommendations

- Several large-scale studies have employed observational measures of parenting rated during a free play task; generally, strong associations have been found between parenting and child outcomes
- It is difficult to assess how similar the scales used in these different studies are and whether they are measuring the same parenting and child constructs
- Methodology papers using the Early Head Start parenting and child observational measures should broaden our knowledge regarding the validity of these scales

IV. Adaptations of Measure

Adaptation:

- N/A

Description of Adaptation

Psychometrics of Adaptation

Study Using Adaptation

Early Childhood Workshop

Early Childhood Workshop

Nursing Child Assessment Satellite Training (NCAST): Teaching Task Scales

Prepared by Allison Sidle Fuligni and Christy Brady-Smith
National Center for Children and Families, Teachers College, Columbia University

I. Background Information

Author/Source

Barnard, K. (1994). *NCAST Teaching Scale*. Seattle, WA: University of Washington, School of Nursing.

Coding and training manual:
Author: Sumner, G., & Speitz, A (1994). *NCAST Caregiver/Parent-Child Interaction Teaching Manual.* Seattle, WA: NCAST Publications, University of Washington, School of Nursing.

Purpose of Measure

- Assess caregiver-child interactions during a semi-structured teaching situation in the home or child care setting
- Utilize discrete, yes/no coding of observable behaviors of both the child and the adult and their interaction

Population Measure Developed With

- The measure has been used in a variety of settings, and the developers have published psychometric information on a large reliability sample (the "NCAST database"; see Sumner & Speitz, 1994).
- The NCAST database includes over 2,000 mother/infant dyads.
- Demographic makeup of the database is 54% Caucasian, 27% African-American, and 19% Hispanic; 77% married mothers; average mother age at child's birth = 25.7 years; average child age 15.5 months (ranging from 0 to 36 months).

Age Range Intended For

- Birth to 36-month-old children with a parent or caregiver
- In Early Head Start, children were assessed at age 24 months

Key Constructs of Measure

- *Parent Constructs*:
  - Sensitivity to cues (caregiver's sensitive responses to child's cues)
  - Response to Child's Distress (caregiver's change of the task and/or comforting responses to a child exhibiting disengagement or distress)

- Social-Emotional Growth Fostering (positive affect and avoidance of negative responses to the child)
- Cognitive Growth Fostering (caregiver's instruction and modeling of the task).

- *Child Constructs*:
  - Clarity of Cues (facial expressions and motor activity indicating child's response to the task situation)
  - Responsiveness to Caregiver (child's facial expressions, vocalizations, and other responses to caregiver)

- In addition to the seven Parent and Child Constructs, there are 4 "Total" scales that may be computed:
  - Parent Total (total of all parent items)
  - Child Total (total of all child items)
  - Caregiver/Child Total (total of all items)
  - Contingency (total of all items, across parent and child constructs, that require behavior that is contingent upon the behavior of the other member of the dyad)

Norming of Measure (Criterion or Norm Referenced)
- The best source for this information is the Sumner & Speitz manual (1994).

Concerns, Comments, and Recommendations:

II. Administration of Measure

Who is the Respondent to the Measure?
- A child, aged 0-36 months, and his or her caregiver or parent.

If Child is Respondent, What is Child Asked to Do?
- The parent or caregiver is asked to select a task that the child can not do. Parents are instructed to explain the task to the child and give the child any necessary assistance in doing the task.
- In the Early Head Start administration, the choice of tasks was limited to two: either sorting blocks or reading a picture book. The interaction lasted three minutes.

Who Administers Measure/ Training Required?

Test Administration
- The protocol for the Early Head Start study was administered by trained interviewer/assessors (I/A)

- I/As were also responsible for videotaping the dyad
- Training sessions for I/As were held at Mathematica Policy Research, Inc. (MPR) and conducted by MPR staff

Data Coding

- Videotapes were coded by a coding team at Columbia University, consisting of 5 coders trained by a certified NCAST instructor during a three-day training course. Each coder was required to pass the NCAST certification in the weeks following the initial training.
- Inter-rater reliabilities between a certified coding team leader and the NCAST-certified coding team were established to a criterion of 85% (exact agreement) on the individual items from the 6 NCAST subscales.
- Intermittent inter-rater reliability checks on a randomly-selected 15% of each coder's videotape assignment were conducted.
- Coders were ethnically heterogeneous
- Interactions conducted in Spanish were rated by a fluent Spanish-speaking coder
- Coders were unaware of participants' treatment group status

Setting (e.g. one on one, group, etc):

- Child-parent interactions were videotaped in the home

Time Needed and Cost

- 4 to 5 minutes per tape for coding
- Estimated cost of graduate student training and videotape coding is $95.00 per videotaped interaction; The NCAST center may be able to provide information on hiring trained NCAST coders to rate videotaped interactions which may significantly lower the cost of coding

Concerns, Comments, and Recommendations:

- In Early Head Start, there were differential rates of children becoming disengaged during the interaction, depending on which task the parent had chosen. Children were more likely to display "potent disengagement cues" when they were engaging in the block-sorting task than in the book-reading task.
- Note that the Early Head Start administration differs in several ways from that described in the training manual (this is discussed further under Adaptations of Measure, below).
- Although the Teaching Scales are designed to be used with children up to age 36 months, many of the coded child behaviors are less relevant to older toddlers. For the Early Head Start administration, developer Kathryn Barnard suggested focusing data analysis on the coded parent items more than the child items.

III. Functioning of Measure

Reliability

Coder reliability: In the Early Head Start study, a total of 130 tapes (8% of the 1687 codable tapes) served as reliability tapes. Percent agreement (exact) on the 6 NCAST subscales ranged from 84% to 95% (*M* = 89%).

Internal reliability: Preliminary analyses of the internal consistency of these scales revealed that very few of the subscales had internal consistency that met the Early Head Start criterion for use as outcome variables in the analyses of program impacts (= .65 or greater). Alphas for the parent subscales ranged from .24 to .74.

- The published psychometric data on the NCAST subscales (Sumner & Speitz, 1994) report internal reliabilities for these subscales that are somewhat higher than those found in the Early Head Start sample. Alphas for the parent subscales reported for the NCAST database range from .52 to .80.
- In the Early Head Start administration, the variability on the Total Score was substantially lower than that found in the NCAST database (K. Barnard, personal communication, 4/11/2001)

Validity

- Sumner & Speitz (1994) report correlations in the NCAST database between parent total scores on this measure and concurrent total HOME scores (Caldwell & Bradley) ranging from .46 to .61, depending on the age of the child. They also report correlations between parent total scores on the Teaching Scales and Bayley Mental Development Index scores of .46 (for a small sample; *N* = 49).
- NCAST teaching scale scores measured at 10 months of age are significantly correlated with 24-month Bayley MDI scores (r =.37, p <.01 for the Parent Total score; Sumner & Speitz, 1994).

Examples of Studies Examining Measure in Relation to Environmental Variation (specify if intervention)

- Such analyses are planned by the members of the Early Head Start Research Consortium (Early Head Start is an intervention)
- See Sumner & Speitz (1994) for more information on such studies.

Concerns, Comments & Recommendations

- A fair amount of psychometric work was conducted with the Early Head Start sample. However, low variability in this sample resulted in low internal reliability for the published subscales. Factor analyses failed to identify a different factor structure for these data, so only the Total and Parent Total scores (= .66 for both scales) were used in Early Head Start impact analyses.

IV. Adaptations of Measure

Adaptation:

- The Early Head Start Research and Evaluation Project adapted the NCAST Teaching Scales for use in a large program evaluation study.

Description of Adaptation

- The procedure was simplified to include only two choices of activity.
- The interaction time was shortened to 3 minutes.
- Coding was done via fine-grained analysis of the videotaped interaction, rather than live, immediately following the observation. This approach allowed for a single group of trained coders to analyze videotapes from all 17 sites of the Early Head Start study, and enabled coders to conduct several passes through the videotape to code the observed behaviors.

Psychometrics of Adaptation

- Reported above; generally, the Early Head Start sample had lower variability than has been found in other samples.
- This may be due to the adaptation of the administration, the homogeneity of the sample (i.e., low-income, low education), and/or the more fine-grained coding that videotaping affords.

Study Using Adaptation

- Early Head Start Research and Evaluation Project