EARLY CHILDHOOD MEASURES PROFILES

Prepared by Child Trends
4301 Connecticut Avenue, NW, Suite 100
Washington, DC 20008
www.childtrends.org

Project Coordinators

Daniel J. Berry
Lisa J. Bridges
Martha J. Zaslow

Authors of Early Childhood Measures Profiles: Lisa J. Bridges, Daniel J. Berry, Rosalind Johnson, Julia Calkins, Nancy Geyelin Margie, Stephanie W. Cochran, Thomson J. Ling, & Martha J. Zaslow; Child Trends.

Authors of Early Head Start Measures section: Allison Sidle Fuligni and Christy Brady-Smith. Center for Children and Families, Teachers College, Columbia University.

This project was made possible by support from the SEED Consortium of federal agencies to the National Institute of Child Health and Human Development Family and Child Well-being Research Network (Grant 1U101 HD 37558-01). The SEED Consortium (Science and the Ecology of Early Development) involves federal agencies working together to lay the groundwork for and fund research on issues pertaining to the development of young children, and especially how specific environments (ecologies) can best support early development. The SEED Consortium agencies that sponsored the present project include the Office of the Assistant Secretary for Planning and Evaluation of the U.S. Department of Health and Human Services, the National Institute of Child Health and Human Development, and the Office of Planning, Research and Evaluation of the Administration for Children and Families of the U.S. Department of Health and Human Services, including the Head Start Bureau and the Child Care Bureau.

The authors thank the following Child Trends staff members for their input into and work on this project: Jacinta Bronte-Tinkew, Kevin Cleveland, Michelle McNamara, LaShaunda Gayden and Laura Wandner.

Table of Contents

Approaches to Learning Measures

Adapted EZ-Yale Personality/Motivation Questionnaire (Adapted EZPQ)

I. Background Information

Author/Source

Source: Wheeler, C., & Henrich, C. (2003). *Measuring the motivation of Head Start students: Validating the EZ-Yale Personality/Motivation Questionnaire.* Poster presented at the Biennial Meeting of the Society for Research in Child Development, Tampa, FL, April.

Publisher: Unpublished. Contact author Christopher Henrich at chenrich@gsu.edu.

Purpose of Measure

As described by the authors

This measure is an adaptation of the EZ-Yale Personality-Motivation Questionnaire (EZPQ; Zigler, Bennett-Gates, Hodapp, & Henrich, 2002). Originally designed to assess academic and social motivation in children ages 5 and older with cultural-familial mental retardation, this adaptation was designed for use with Head Start children. According to Wheeler and Henrich (2003), "Prior work with older children has indicated that children in poverty experience similar motivational challenges to those with mental retardation."

Population Measure Developed With

- The sample included 133 4 year-old children enrolled in Head Start in Austin, Texas. Sixty-two percent of the children were female, 65 percent were black, 32 percent were Latino, and 3 percent were white. The median household income for the sample was $5,819. Half of the children were enrolled in their first year of Head Start, half were in their second year of enrollment (i.e., they began Head Start at age 3).
- Head Start teachers completed the original 37-item EZPQ for children in their classrooms as part of a larger school readiness assessment.

Age Range Intended For

Children enrolled in preschool. The sample for which the measure was developed included only 4 year-olds.

Key Constructs of Measure

The adapted EZPQ includes three scales. In total, 24 of the original 37 EQPQ measures are included.

- *Academic Mastery Motivation*. This scale includes 12 items from three scales of the original EZPQ: Effectance Motivation, Creativity/Curiosity, and Expectancy of Success. Items tap persistence, initiative, creativity, the tendency to work for pleasure or without expectation of a tangible reward, and expectations that efforts will be successful.
- *Negative Reaction Tendency*. This scale consists of all six items from the original EZPQ Negative Reaction Tendency scale. High scores reflect social withdrawal.
- *Outerdirectedness*. The six-item Outerdirectedness scale from the EZPQ was retained. High scores on this scale indicate the tendency to imitate others and to be a follower, rather than a leader of social activities.

Norming of Measure (Criterion or Norm Referenced)
No norming information is currently available.

Comments
The adapted EZPQ is a new measure that does not yet have a track record for assessment of young children's academic and social motivation. Because of the relative lack of measures designed to assess academic mastery motivation in young children, particularly measures that could be useful in large-scale basic and evaluation research, this measure may be a promising step forward. Future work is necessary in order to determine its value, however. Of particular importance will be additional work with larger samples that are representative of preschool children in general or more specifically of Head Start children that can be used to provide norms for this measure and to establish thresholds that may reflect particularly high or low levels of adaptation to preschool settings.

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 adapted EZPQ is a teacher-report questionnaire that is completed independently by respondents. As such, little training is required for test administration.

Data Interpretation
Currently, the adapted EZPQ has been used only for research purposes, rather than for evaluations of individual children or programs. Scoring is objective and requires little interpretation when used as a research instrument.

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

Time Needed and Cost
Time
No time estimate is provided but should require no more than 5 to 10 minutes per child.

Cost
There is no cost at this time. The measure is not published.

III. Functioning of Measure

Reliability Information from Manual

Internal Consistency Reliability

Cronbach's alphas for Academic Mastery Motivation, Negative Reaction Tendency, and Outerdirectedness were .92, .72, and .71, respectively.

Validity Information from Manual

Criterion Validity

Wheeler and Henrich (2003) examined criterion validity by correlating scores on the three EZPQ scales with four measures of school readiness in the academic domain and five measures of school readiness in the social competence domain. The four academic readiness measures were the Letter-Word Identification, Dictation, and Math subtests from the Woodcock Johnson (WJ, version not specified; see McGrew & Woodcock, 2001), and the Peabody Picture Vocabulary Test III (PPVT-III; Dunn & Dunn, 1997). Wheeler and Henrich report partial correlations for each EZPQ scale score with each of the school readiness measures, controlling for scores on the other two EZPQ scale scores. The social competence readiness scores included five scales from the Social Skills Rating System (SSRS; Gresham & Elliott, 1990), including Cooperation, Assertion, Self-Control, Internalizing, and Externalizing.

- Academic Mastery Motivation was significantly correlated with all criterion measures. Controlling for Negative Reaction Tendency and Outerdirectedness scores, Academic Mastery Motivation scores had correlations ranging from .26 to .37 with academic school readiness measures. Partial correlations with the social competence school readiness scores ranged from -.30 with both SSRS Internalizing and Externalizing scores to .51 with SSRS Cooperation.
- One of four partial correlations between Negative Reaction Tendency and academic school readiness measures was significant. Controlling for Academic Mastery Motivation and Outerdirectedness, Negative Reaction Tendency was correlated -.21 with PPVT-III scores. The remaining three correlations ranged from -.05 to -.11. Of the five partial correlations with social competence measures, two were significant: The partial correlation with SSRS Assertion was -.36, while the correlation with SSRS Internalizing was .31. Other partial correlations ranged from .03 to -.14.
- Outerdirectedness was correlated -.18 with WJ Math scores. Other correlations with academic readiness measures were not significant, ranging from -.07 to -.13. Three of five correlations with social competence school readiness measures were significant. EZPQ Outerdirectedness correlated -.26 with SSRS Cooperation, -.22 with SSRS Self-Control, and .26 with SSRS Externalizing. The remaining two correlations, with SSRS Assertion and SSRS Internalizing, were -.15 and .06, respectively.

Comments

- The Academic Mastery Scale of the EZPQ is the most pertinent component of this measure for the examination of approaches to learning. This scale showed consistent correlations with academic and social aspects of school readiness, although correlations were not strong.
- The sample used for the development of this measure was primarily composed of ethnic minority (black or Latino) children from low-income families, and all of the children

were participating in a Head Start program. Additional work with larger and more representative samples will be needed in order to determine the usefulness of this measure for assessment of individual children or for program evaluation.

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

Wheeler and Henrich (2003) included comparisons of adapted EZPQ scores of children who were enrolled in their first year of Head Start, versus children who had been enrolled since age 3, in order to determine "…whether starting Head Start at age 3 was associated with a more adaptive motivational profile." Consistent with expectations, children in their second year of Head Start had significantly higher Academic Mastery Motivation scores and significantly lower Negative Reaction Tendency and Outerdirectedness scores than did children in their first year of Head Start participation.

V. Adaptations of Measure

None found. This measure is an adaptation of a measure originally developed for use with older children and young adults with mental retardation.

ECLS-K[1] 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.

[1] 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.

Data Interpretation
This measure has not been used for evaluations of individual children or programs. Scoring is objective and requires little interpretation when used as a research instrument.

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 assessment 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.

Games as Measurement for Early Self-Control (GAMES)

I. Background Information

Author/Source

Source: McCabe, L. A., Hernandez, M., Lara, S. L., & Brooks-Gunn, J. (2000). Assessing preschoolers' self-regulation in homes and classrooms: Lessons from the field. *Behavioral Disorders, 26*, 42-52.

McCabe, L. A., Rebello-Britto, P., Hernandez, M., & Brooks-Gunn, J. (2004). Games children play: Observing young children's self-regulation across laboratory, home, and school settings. In R. DelCarmen-Wiggins & A. Carter (Eds.), *Handbook of infant, toddler, and preschool mental health assessment*. New York: Oxford University Press.

Publisher: Unpublished. Contact author Lisa McCabe at lm428@columbia.edu.

Purpose of Measure

As described by the authors

The goal of the Games as Measurement for Early Self-Control (GAMES) project was to "pilot, modify, and pilot again self-regulatory tasks used in laboratory-based studies for use in the home and preschool, keeping in mind the need for relatively simple-to-administer tasks (i.e., those that do not require excessive staff training to reach reliability) that are appropriate for many preschoolers" (McCabe, Hernandez, Lara, & Brooks-Gunn, 2000, p. 54).

Population Measure Developed With

- The initial sample described by McCabe *et al*. (2000) included 71 3- to 5-year-old children (34 boys) from predominantly low-income families, with incomes reported as ranging from "…less than $5,000 to more than $60,000" (p. 55). Families were recruited through Head Start and other programs serving low-income families. The sample was predominantly Latino (66 percent); 18 percent of the children were black, 6 percent were white, and 10 percent were from other ethnic and racial backgrounds (including multiracial). Both English- and Spanish-speaking children were included in the sample.
- In a later report (McCabe, Rebello-Britto, Hernandez, & Brooks-Gunn, 2004), the full sample included 116 children. Compared with the smaller samples described by McCabe *et al*. (2000), a somewhat lower percentage (54 percent) was Hispanic and a somewhat higher percentage (29 percent) was black.
- McCabe *et al*. (2004) describe additional piloting of self-regulation situations similar or the same as those included in the GAMES project conducted in the Storytimes Study, involving 40 preschool children enrolled in the Home Instruction Program for Preschool Youngsters in Bronx and Yonkers, New York (Baker, Piotrkowski, & Brooks-Gunn, 1998). These children were from low-income families; their mean age was 4.37 years (SD = .34 years) and 45 percent were boys.

Age Range Intended For

Ages 3 to 5.

Key Constructs of Measure
The tasks included in GAMES are largely adaptations of procedures designed for laboratory-based studies of effortful control and self-regulation in young children (e.g., Kochanska, Murray, & Coy, 1997; Kochanska, Murray, Jacques, Koenig, & Vandegeest, 1996; Maccoby, Dowley, Hagen, & Degerman, 1965); some tasks were created specifically for home- and school-based assessments. Tasks were adapted or created in order to examine several abilities related to self-regulation in young children.

- *Motor Control*. The ability to control both fine and gross motor movements. Examples of motor control tasks include drawing a circle in between a larger and a smaller circle at varying speeds, and walking along a line at varying speeds.
- *Impulse Control/Delay of Gratification.* The ability to inhibit behavior in order to achieve a desired goal. Examples of tasks include waiting for a snack, or waiting without peeking while a researcher "wraps" a gift that will subsequently be given to the child to open.
- *Cognitive Control.* The ability to inhibit a dominant response in favor of a less common response. Examples of tasks include a version of "Simon Says" using hand puppets in which children are told to do what one puppet says but not to do what the second puppet says, and a game in which children are told to touch their heads when the researcher says "feet," and vice versa.
- *Sustained Attention.* The ability to sustain focused attention on a task. Examples of tasks include asking a child to make a card for someone while the data collector filled out forms, and mother-child structured play situations.

Norming of Measure (Criterion or Norm Referenced)
No norming information is currently available.

Comments
GAMES represents an interesting effort to develop measures of self-regulation abilities, many adapted from tasks that were originally designed for laboratory and clinical settings, in order to examine these abilities through assessments in the home and other real-life settings. Because no norms have yet been established for individual tasks or for a standardized GAMES battery of tasks, GAMES cannot be used for assessment of individual children's functioning. Much work remains to be done in order to determine whether GAMES measures can be standardized and normed so as to be useful in evaluations of programs or individual children.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
There are a variety of tasks included in GAMES, all of which are designed to be challenging, but at the same time fun and interesting, for the child. Examples include walking along a line at

varying speeds, drawing circles at varying speeds, and waiting without peeking for an adult to "wrap" a gift that will subsequently be given to the child to open.

Who Administers Measure/Training Required?

Test Administration

No specific qualifications were described by McCabe *et al*. (2000) for individuals administering GAMES tasks. Administrators need to be well-trained in the procedures and able to make acceptable adaptations depending upon testing conditions and cultural considerations, without making adaptations that would undermine the ability to interpret results. Some GAMES procedures can be used with live coding, while others require videotaping and subsequent coding from tapes.

Data Interpretation

Currently, GAMES measures have been used only for research purposes, rather than for evaluations of individual children or programs. Interpretation of scores has generally involved examining associations between GAMES scores and scores on other child, parent, or environmental measures; requirements for interpretation would thus involve training in conducting and interpreting results of statistical analyses commonly used in psychological and educational research.

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

This assessment is usually administered in a one-on-one setting. Some procedures have also been used with small groups of preschoolers in a classroom setting.

Time Needed and Cost

Time

No specific time estimates are given. GAMES procedures are relatively short, lasting a few minutes each. McCabe *et al.* (2004) indicated that children in the GAMES development project took approximately 30 minutes to complete between six and eight tasks.

Cost

There is no cost at this time. The measures are not published.

III. Functioning of Measure

Reliability Information from Manual

No specific reliability information was provided by the authors.

Validity Information from Manual

No specific validity information was provided by the authors.

Reliability/Validity Information from Other Studies

Pittman, Li-Grining, and Chase-Lansdale (2002) used data from the Embedded Developmental Study (EDS), a sub-study of low-income families living in high-poverty neighborhoods who participated in Welfare, Children, and Families: A Three City Study. Approximately 580

mothers with children between the ages of 2 and 4 agreed to participate; families were predominantly Hispanic (51 percent) and black (43 percent). During home visits, mother-child interactions and child self-regulation tasks were administered and videotaped. Two self-regulation tasks similar to those examined in the GAMES project were analyzed: 1) a Snack Delay in which the child had to wait varying lengths of time before picking up and eating a piece of candy, and 2) Gift Wrap, during which children were asked to face away and not peek while the investigator wrapped a gift that she had brought for the child.

Interrater Reliability
Seven coders rated children's self-regulation from videotapes. Interrater reliability (kappa) for a 10-point behavioral rating scale used for Snack Delay averaged .69; the average intraclass correlation for a measure of latency to eat was .98. Interrater reliability for the 8-point behavioral rating scale used for the Gift Wrap Task was .62; intraclass correlations for latency to peek and latency to turn toward the investigator averaged .94 and .80, respectively (see Pittman *et al.*, 2002, p. 7).

Concurrent Validity
Pittman *et al.* (2002, p. 20) correlated self-regulation ratings with other child characteristics, including temperamental impulsivity, [negative] emotionality, and sociability as rated by mothers, and noncompliance, persistence, and negative affect during a challenging structured mother-child play interaction. Correlations between a total self-regulation measure created by compositing Snack Delay and Gift Wrap and behaviors during the structured play situation were low to moderate and significant, ranging from -.22 for negative affect to .38 for persistence. Correlations with mother-rated temperament characteristics were lower, ranging from nonsignificant correlations of .07 and -.07 with emotionality and sociability, respectively, to a significant correlation of -.15 with impulsivity.

Comments

- In selecting measures to be included in a GAMES battery, the authors paid close attention to issues related to reliability and validity, including standardized administration, variability in children's performance, whether performance differences were truly associated with self-regulation abilities versus confusion regarding task instructions, cultural sensitivity and appropriateness, and whether tasks could be coded reliably and appropriately, either during live coding or from videotapes. However, little specific reliability or validity information is currently available. Although developments at this time are promising, much work remains to be conducted in order to determine whether observational measures of self-regulation have sufficient reliability and validity to be used for evaluations of individual children and programs.
- Some limited support for both reliability and validity of self-regulation measures such as those included in GAMES is provided by results presented by Pittman *et al.* (2002). For interrater reliability of the rating scales, only average kappas were reported, and it is thus not clear what the full range of kappas was in this study. Kappas are a measure of exact agreement between raters. Although the reported average kappas, .69 for a 10-point scale and .62 for an 8-point scale, are generally seen as acceptable in observational research with young children, they also indicate that raters were far from being in perfect agreement regarding the assignment of exact codes. To the extent that these scales are

ordinal (i.e., to the extent that codes are arranged on a continuum from lesser to greater self-regulation), exact agreement may be a somewhat conservative method for determining interrater reliability; intraclass correlations between ratings provided by different coders would be helpful in further clarifying this issue. In general, it may be that further refinement of measurement scales are necessary before ratings of behavior using methods such as GAMES can be used reliably for individual child or program assessment.

- With respect to concurrent validity, the composite self-regulation measures had expected associations with other child measures, including persistence, compliance, affect during challenging task situations, and impulsivity. Although significant, however, these correlations were low to moderate in magnitude. Additional research using other outcome measures that are hypothesized to be associated with children's ability to self-regulate their behavior will be needed to determine the extent to which GAMES can be used reliably to assess such ability.

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

Numerous laboratory-based studies with relatively small, non-representative samples, using tasks similar to those modified or created for the GAMES project, have been published. In addition, a number of large scale, multi site studies are currently underway that include one or more tasks such as those developed in the GAMES project. These studies include the Early Head Start Research and Evaluation Project, the Fragile Families and Child Well-Being Study, the Project on Human Development in Chicago Neighborhoods, and Welfare, Children, and Families: A Three City Study (see Brooks-Gunn, Berlin, Leventhal, & Fuligni, 2000). At this time, no reports including self-regulation data from GAMES-type tasks administered in home or classroom settings, or the effects of environmental variations on self-regulation abilities, have been published based on these studies (the Pittman *et al.* results summarized above were presented at a 2002 workshop on new directions in socio-emotional measures).

V. Adaptations of Measure

None found. GAMES measures are evolving, and many GAMES tasks are themselves adaptations of measures originally developed for laboratory-based studies by other authors.

NEPSY

I. Background Information

Author/Source

Source: Korkman, M., Kirk, U., & Kemp, S. (1998). *NEPSY: A Developmental Neuropsychological Assessment. 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 the authors

"The NEPSY is a comprehensive instrument that was designed to assess neuropsychological development in preschool and school-age children. The name, NEPSY, is an acronym that was formed from the word neuropsychology, taking NE from *neuro* and PSY from *psychology*" (Korkman, Kirk, & Kemp, 1998, p. 1). Identified uses for the NEPSY include assessments of children at risk for later problems due to a variety of conditions (e.g., brain damage, very low birth weight, lead exposure), long-term follow-up of at-risk children, identification of more subtle problems that may interfere with learning, and the study of both typical and atypical neuropsychological development.

Population Measure Developed With

- The standardization sample for the NEPSY included 1,000 children ages 3 years, 0 months through 12 years, 11 months. There were 100 children in each 1-year age group, 50 males and 50 females. The median age for each group was the fifth month (e.g., 3 years, 5 months; 12 years, 5 months).
- In order to achieve a nationally-representative sample, the sample was stratified by race/ethnicity (White, African American, Hispanic, and Other), geographic region (Northeast, North Central, South, and West), and parent education (11th grade or less, high school graduate or equivalent through 3 years of college, and 4 or more years of college). A total of 200 examiners provided assessments of children. These examiners were selected on the basis of assessment experience, location, and the ages, socioeconomic status, and race/ethnicity of the children to whom they had access.
    - Approximately 16 percent of the resulting sample were African American, 12 percent were Hispanic, 69 percent were White, and 4 percent were Other (see Korkman *et al.*, 1998, p. 33)
    - Geographically, approximately 20 percent of the sample was drawn from the Northeast, 23 percent were from the North Central region, 35 percent were from the South, and 22 percent were from the West.

- Ten percent of parents had an eleventh grade education or less, 60 percent had at least a high school diploma or equivalent but less than 4 years of college, and 30 percent had 4 or more years of college.

Age Range Intended For
Ages 3 years, 0 months through 12 years, 11 months.

Key Constructs of Measure
There are five domains covered by the NEPSY.

- *Attention/Executive Functions*: Executive function involves "…planning and flexible strategy employment…; the ability to adopt, maintain, and shift cognitive set, to use organized search strategies, and to monitor performance and correct errors; the ability to resist or inhibit the impulse to respond to salient, but irrelevant aspects of a task…; and working memory" (Korkman *et al.*, 1998, p. 11). The subtests included in this domain tap abilities associated with self-regulation of auditory and visual attention and behavior.
- *Language:* Language subtests tap language abilities related to phonological processing, naming, receptive language comprehension, and language production.
- *Sensorimotor Functions:* Subtests examine fine motor skills, eye-hand coordination, the ability to imitate movements and positions, and the ability to correctly identify tactile input without visual input.
- *Visuospatial Processing:* Subtests assess the ability to reproduce two- and three-dimensional figures as well as the understanding of directionality, line orientation, and visuospatial relationships.
- *Memory and Learning:* Subtests tap memory for faces, names, lists, sentences, and story details.

The number of subtests within each domain, and the abilities tapped by these subtests differ for preschool (3- and 4-year-old) and school-age (5- through 12-year-old) children. There are recommended Core Assessments for each age group that include subtests from each domain; Expanded Assessments include all age-appropriate subtests within each domain. In addition, the NEPSY is designed so that domains and subtests can be sampled based on the specific reasons for assessment, as well as constraints such as time and setting in which assessment takes place.

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

Comments
The NEPSY is a highly clinical instrument, the primary purpose of which is to assess children for a variety of neuropsychological problems, included disorders of attention, hyperactivity, language, reading, arithmetic, and motor coordination. The aspect of the NEPSY that may be most relevant to Approaches to Learning is Attention/Executive Functions; the abilities included within this domain are related to children's abilities to maintain focus on tasks, to follow instructions, and to persist in task-related behaviors.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
NEPSY assessments include a variety of tasks in which children are asked to observe the evaluator's actions, to listen to orally presented stimuli, and to follow directions regarding the types of responses requested. Responses for different subtests vary and include oral responses, drawing, writing, working with blocks, puzzles, or other manipulatives, and hand or body movements or movement inhibition.

Who Administers Measure/Training Required?
Test Administration
According to the authors, the NEPSY should be administered by individuals with graduate training in the administration of standardized psychological and cognitive assessments. The authors further indicate, however, that a well-trained research assistant or technician can administer and score NEPSY subtests under the supervision of an examiner with graduate training in assessment. Regardless of training, examiners should have prior experience testing children similar to those being examined with the NEPSY, with respect to characteristics such as age and linguistic, ethnic, cultural, and socioeconomic background.

Data Interpretation
Interpretation of NEPSY profiles requires graduate-level training in psychological assessment; when the NEPSY is to be used as a neuropsychological assessment, specific training in neuropsychology is required.

Setting (e.g. one-on-one, group, etc.)
This test is administered in a one-to-one setting.

Time Needed and Cost
Time
According to the Manual, a NEPSY Core Assessment takes approximately 45 minutes for preschoolers, approximately one hour for children age 5 and older. If the focus is on Attention/Executive Functions domain subtests, estimated time for preschoolers is approximately 5 minutes for the two subtests, and between 19 minutes (three subtests) and 28 minutes (six subtests) for older children.

Cost
- A complete kit including all materials necessary for administering the NEPSY is $598.
- A kit including scoring assistance software is $650.
- Packets of 25 record forms are $31 for preschool assessments, $37 for assessments of children age 5 and older.
- Response booklets (necessary for some subtests are $28 and $34 for younger and older child assessments, respectively.

III. Functioning of Measure

Reliability Information from Manual

Internal Consistency Reliability

Internal consistency reliability was computed using a split-half procedure in which subtests were divided into two halves of equal lengths and scores for each half were correlated and corrected for length of the test. Internal consistency reliability estimates were derived using this procedure for 14 of the 22 subtests. Characteristics of the remaining subtests, including how they were administered and scored, precluded this procedure for estimating reliability. Correlations are presented by the authors for single-year age groups, and average correlations (calculated using Fischer's *z* transformation) are presented for preschoolers (ages 3 and 4) and for school-age children (ages 5 through 12; see Korkman *et al.*, 1998, p. 177).

- For Attention/Executive Functions, internal consistency was calculated for a single subtest (Tower) which is not administered to preschoolers. At ages 5 and 6 the internal consistencies of this subtest were .89 and .90, respectively, and the average across ages 5 through 12 was .82.
- Internal consistencies were available for three Language subtests for preschoolers (Body Part Naming, Phonological Processing, and Comprehension of Instructions), and for three subtests for school-age children (Phonological Processing, Comprehension of Instruction, and Repetition of Nonsense Words).
    - At ages 3 and 4, correlations ranged from .73 (Body Part Naming at age 3) to .89 (Comprehension of Instructions at age 4), and preschool average correlations ranged from .74 (Body Part Naming) to .89 (Comprehension of Instructions).
    - At ages 5 and 6, correlations ranged from .82 (Repetition of Nonsense Words at age 5) to .93 (Phonological Processing at age 6). School-age averages ranged from .73 (Comprehension of Instructions) to .91 (Phonological Processing).
- Internal consistency coefficients were calculated for one Sensorimotor subtest at all ages (Imitating Hand Positions). Between ages 3 and 6, correlations ranged from .88 at age 4 to .91 at age 5. The preschool average reliability was .89, while the school-age average was .82.
- All Visuospatial subtests—Design Copying, Block Construction, and Arrows (ages 5 and older)—could be examined for internal consistency.
    - At ages 3 and 4, correlations ranged from .80 (Block Construction at both ages) to .87 (Design Copying at age 4). Preschool average correlations were .86 for Design Copying and .80 for Block Construction.
    - At ages 5 and 6, correlations ranged from .69 (Block Construction at age 5) to .88 (Arrows at age 6). School-age average correlations ranged from .72 (Block Construction) to .79 (Design Copying).
- Internal consistencies could be estimated for all Memory and Learning subtests. Two of these subtests (Narrative Memory and Sentence Repetition) are used with children of all ages; two subtests are used with children ages 5 and older (Memory for Faces and Memory for Names); one additional subtest is used only with children ages 7 and older (List Learning).

- Correlations at ages 3 and 4 ranged from .84 (Narrative Memory at age 4) to .91 (Sentence Repetition at both ages). Average preschool correlations were .85 for Narrative Memory and .91 for Sentence Repetition.
- Correlations at ages 5 and 6 ranged from .74 (Memory for Faces at age 6) to .91 (Memory for Names at age 6). Average school-age correlations ranged from .76 (Memory for Faces) to .91 (List Learning).

Test-Retest Stability

Test-retest data were presented by Korkman *et al.* (1998, pp. 181-185) for all subtests. Stability coefficients (Pearson correlations corrected for variability within the full standardization sample) were provided for most subtests. Due to highly skewed distributions and limited variability, the stability statistic reported for some subtests was the percentage of children who remained within the same classification range (less than or equal to the 10th percentile, 11th to 75th percentile, or greater than 75th percentile) at both assessments. All test-retest data were collected for a sample of 168 children (49 percent male) to whom the NEPSY was administered twice. Time between testing sessions ranged from 2 to 10 weeks, with an average interval of 38 days. Test-retest information was provided for 3- to 4-year-olds ($N = 30$), 5- and 6-year-olds ($N = 33$), 7- and 8-year-olds ($N = 31$), 9- and 10-year-olds ($N = 41$), and 11- and 12-year-olds ($N = 33$), as well as averaged correlations across the full school-age range (5 to 12 years).

- Attention/Executive Functions.
  - For 3- and 4-year-olds, the stability coefficient for standardized Attention/Executive Functions domain scores was .68; coefficients for subtests were .69 and .50 for Visual Attention and Statue, respectively.
  - In the 5- and 6-year-old age group, the stability coefficient for Attention/Executive Functions domain scores was .80. Test-retest correlations for 4 subtests (Tower, Auditory Attention and Response Set, Visual Attention, and Design Fluency) ranged from .63 (Design Fluency) to .84 (Auditory Attention and Response Set). For two additional subtests with skewed distributions, Statue and Knock and Tap, percentages of children whose scores fell within the same range at both assessments were 75 percent and 63 percent, respectively.
  - The average stability coefficient for domain scores across the school-age range was .67, while the average correlations for subtests ranged from .53 (Tower) to .81 (Auditory Attention and Response Set). Average percentages of children whose scores for Statue and Knock and Tap fell within the same range were 69 percent and 65 percent, respectively.
- Language.
  - In the 3- and 4-year-old age group, the stability coefficient for Language domain scores was .78; correlations for four subtests (Body Part Naming, Phonological Processing, Comprehension of Instructions, and Verbal Fluency) ranged from .42 (Phonological Processing) to .65 (Comprehension of Instructions), and the percentage of children whose scores fell within the same range across the two assessments for one additional subtest, Oromotor Sequences, was 50 percent.
  - In the 5- and 6-year-old group, the stability coefficient for Language domain scores was .78. Correlations for five subtests (Phonological Processing, Speeded Naming, Comprehension of Instructions, Repetition of Nonsense Words, and Verbal Fluency) ranged from .53 (Phonological Processing) to .86 (Repetition of

Nonsense Words). On the Oromotor Sequencing subtest, 67 percent of children had scores that fell within the same classification range.
    - Across the full 5- to 12-year-old age range, the average stability coefficient for Language domain scores was .76; average stability coefficients for subtests ranged from .55 (Comprehension of Instructions) to .76 (Repetition of Nonsense Words), and the average percentage of children whose scores for Oromotor Sequences fell within the same range at both assessments was 62 percent.
- Sensorimotor.
    - The stability coefficient for Sensorimotor domain scores for 3- and 4-year-olds was .77. Subtest test-retest correlations were .62 (Visuomotor Precision) and .77 (Imitating Hand Positions); for a third subtest, Manual Motor Sequences, 47 percent of children scored within the same range at both assessments.
    - In the 5- and 6-year-old age group, the test-retest correlation for Sensorimotor domain scores was .81. Correlations for 3 subtests (Fingertip Tapping, Imitating Hand Positions, and Visuomotor Precision) ranged from .66 (Imitating Hand Positions) to .78 (Visuomotor Precision). For three additional subtests with skewed distributions, Manual Motor Sequences, Finger Disc-Preferred, and Finger Disc-Nonpreferred, percentages of children whose scores fell within the same range at both assessments were 63, 55, and 45 percent, respectively.
    - Across the school-age range, the average stability coefficient for Sensorimotor domain scores was .67; average stability coefficients for subtests ranged from .53 (Imitating Hand Positions) to .71 (Fingertip Tapping). For Manual Motor Sequences, Finger Disc-Preferred, and Finger Disc-Nonpreferred subtests, 54, 61, and 56 percent of children had scores within the same classification range at both assessments.
- Visuospatial.
    - In the 3- and 4-year-old age group, the stability coefficient for Visuospatial domain scores was .72; correlations for two subtests, Design Copying and Block Construction, were .71 and .56, respectively.
    - For 5- and 6-year-olds, the stability coefficient for Visuospatial domain scores was .79. Test-retest correlations for 3 subtests (Design Copying, Arrows, and Block Construction) ranged from .59 (Arrows) to .81 (Design Copying). On the Route Finding subtest, 58 percent of children had scores that fell within the same classification range at both assessments.
    - Across the 5- to 12-year-old age range, the average stability coefficient for Visuospatial domain scores was .70. Average stability coefficients for subtests ranged from .52 (Arrows) to .74 (Design Copying). The average percentage of children whose scores for the Route Finding subtest fell within the same range at both assessments was 65 percent.
- Memory and Learning.
    - The stability coefficient for Memory and Learning domain scores for 3- and 4-year-olds was .90. Correlations were .81 and .89 for Narrative Memory and Sentence Repetition subtests, respectively.
    - In the 5- and 6-year-old age group, the stability coefficient for Memory and Learning domain scores was .83; corrected test-retest correlations for the 4 subtests (Memory for Faces, Memory for Names, Narrative Memory, and

Sentence Repetition) ranged from .70 (Memory for Faces) to .78 (Narrative Memory).
- For age 5 through 12, the average stability coefficient for standardized domain scores was .76, and coefficients for subtests ranged from .57 (Memory for Faces) to .76 (Sentence Repetition).

Generalizability
In addition to test-retest information, generalizability coefficients were calculated as a measure of reliability for three subtests (two for preschoolers) that involve both speed and accuracy components. According to Korkman *et al.* (1998, p. 176), this coefficient "…was calculated to account for the multiple sources of error that are present due to the multidimensional nature of these tasks…" Of these subtests, one (Visual Attention) falls within the Attention/Executive Functions domain for children ages 5 and older, one (Speeded Naming) is a Language subtest, and one (Visuomotor Precision) is a Sensorimotor subtest.

- Visual Attention generalizability coefficients were .76 for ages 3 and 4, and .68 for ages 5 and 6. The average generalizability coefficient for ages 5 through 12 was .71.
- Visuomotor Precision generalizability coefficients were .81 at ages 3 and 4, and .88 at ages 5 and 6. The average coefficient for ages 5 through 12 was 68.
- The Speeded Naming generalizability coefficient was .73 for ages 5 and 6, and the average coefficient for ages 5 through 12 was .74.

Interrater Reliability
Interrater reliability was assessed with a subsample of 50 randomly selected cases from the standardization sample that were scored independently by two trained raters. Only three subtests were included in this study – one Visuospatial subtest (Design Copying), one Sensorimotor subtest (Visuomotor Precision), and one Language subtest for children ages 5 and older (Repetition of Nonsense Words). These subtests were included because scoring of each requires some subjective decision-making on the part of the rater. Intraclass correlations, adjusted for scorer leniency, were calculated. For the three tests, interrater reliabilities ranged from .97 to .99. No information on interrater reliabilities for different child age groups was provided (see Korkman *et al.*, 1998, p. 181).

Validity Information from Manual
Content Validity
Contents of the NEPSY subtests were initially based on A. R. Luria's theoretical work regarding neuropsychological development. Other aspects from current cognitive research and psychometric design have been added to improve the ability of the NEPSY to assess children with neuropsychological and developmental disorders. Test content was reviewed by multiple panels of experts, including pediatric neuropsychologists and school psychologists, for both test content and breadth of coverage. Revisions to the NEPSY were made based on reviewer commentary and data from pilot and tryout phases of the test design.

Construct Validity
Construct validity was assessed in a variety of ways. First, correlations among domain scores were examined, with the expectation that correlations would be moderate among the various domain scores within a nonclinical sample of children.

- In the 3- and 4-year-old age range correlations ranged from .34 between Attention/Executive Functions and Language, to .65 between Language and Memory and Learning (see Korkman *et al.*, 1998, p. 361).
- Correlations among domain scores in the 5- to 12-year-old group ranged from .18 (Visuospatial and Memory and Learning) to .46 between Language and Memory and Learning (see Korkman *et al.*, 1998, p. 365).

Construct validity was also examined by focusing on intercorrelations among subtests within and across domains.

- Ages 3 and 4.
    - The two Attention/Executive Functions subtests correlated .24. Correlations among the five Language subtests ranged from .23 to .59 (median = .38). Correlations among the three Sensorimotor subtests were all .25. The two Visuospatial subtests correlated .40, as did the two Memory and Learning subtests.
    - Across-domain median correlations ranged from .23 for Attention/Executive Functions subtests correlated with Language subtests and for Sensorimotor subtests correlated with Memory and Learning subtests, to .45 for Language subtests correlated with Memory and Learning subtests (see Korkman *et al.*, 1998, p. 361).
- Ages 5 through 12.
    - Attention/Executive Functions subtest correlations ranged from .02 to .93; all but 3 correlations were .25 or below and the median correlation was .14. For Language, correlations ranged from .25 to .46 (median = .34). Sensorimotor subtest correlations ranged from .07 to .51 (median = .15); only one correlation was above .25. Memory and Learning subtests correlated between .10 and .96 (median = .27); only six correlations were above .42. Visuospatial subtests correlated between .34 and .44 (median = .36).
    - Across domains, the lowest median correlations were .12 for Attention/ Executive Functions subtests correlated with both Sensorimotor subtests and Memory and Learning subtests, and Sensorimotor subtests correlated with Memory and Learning subtests. The highest median correlation, .27, was between Language subtests and Memory and Learning subtests (see Korkman *et al.*, 1998, pp. 362-365).

Korkman *et al.* (1998) also described a series of validity studies in which NEPSY scores were compared to scores on general cognitive assessments, achievement tests, academic performance, and neuropsychological assessments. The majority of these studies were conducted with school-age children. Only two of the reported studies included preschool-age children.

- In one preschool study, the NEPSY (Core subtests) and the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R; Wechsler, 1989) were administered to a nonclinical sample of 45 children with a mean age of 4.04 years (*SD* = 0.90 years); 56 percent of children in the sample were male (see Korkman *et al.*, 1998, pp. 197, 203).
    - Attention/Executive Functions domain scores correlated .24 with WPPSI-R Verbal IQ scores, .24 with Performance IQ scores, and .26 with Full Scale IQ scores.

  - Language domain scores correlated .60 with Verbal IQ scores, .45 with Performance IQ scores, and .57 with Full Scale IQ scores.
  - Sensorimotor domain scores correlated .34 with Verbal IQ scores, .25 with Performance IQ scores, and .31 with Full Scale IQ scores.
  - Visuospatial domain scores correlated .44 with Verbal IQ scores, .42 with Performance IQ scores, and .47 with Full Scale IQ scores.
  - Memory and Learning domain scores correlated .51 with Verbal IQ scores, .45 with Performance IQ scores, and .51 with Full Scale IQ scores.
- In a second study with young children, associations between NEPSY domain scores and scores on the Bayley Scales of Infant Development-Second Edition (BSID-II; Bayley, 1993) were examined in a nonclinical sample of 20 3-year-olds; 35 percent of children in the sample were male (see Korkman *et al.*, 1998, pp. 197, 203).
  - Attention/Executive Functions domain scores correlated -.31 with BSID-II Mental Index scores and -.37 with Psychomotor.
  - Language domain scores correlated .61 with Mental Index scores and -.11 with Psychomotor Index scores.
  - Sensorimotor domain scores correlated .31 with Mental Index scores and .22 with Psychomotor Index scores.
  - Visuospatial domain scores correlated -.04 with Mental Index scores and -.09 with Psychomotor Index scores.
  - Memory and Learning domain scores correlated .05 with Mental Index scores and -.56 with Psychomotor Index scores.
- Korkman *et al.* (1998) reported a series of studies examining associations between NEPSY domain scores and achievement among school-age children, as measured through grades and standardized achievement tests.
  - Correlations with recent grades in language, mathematics, writing, science, and social science were examined in a sample of 445 children (mean age = 9.52 years, *SD* = 1.99 years, 49 percent male). The lowest correlations with grades were found for Attention/Executive Functions (range = .10 to .17) and Sensorimotor (range = .13 to .17), while the highest correlations were found for Language (range = .33 to .40; see Korkman *et al.*, 1998, pp. 197, 205).
  - Correlations between NEPSY scores and scores on the Wechsler Individual Achievement Test (WIAT, 1992) were reported for a sample of 39 children diagnosed with a learning disability (mean age = 9.46 years, *SD* = 1.59 years, 51 percent male). Correlations for the five NEPSY domain scores and WIAT Total composites were -.02, .14, .16, .21, and .34 for Sensorimotor, Memory and Learning, Attention/Executive Functions, Visuospatial, and Language domains, respectively (see Korkman *et al.*, 1998, pp. 197, 206).
  - Correlations were also reported between NEPSY domain scores and scores on group-administered achievement tests for a nonclinical sample of 304 children (mean age = 9.87 years, *SD* = 1.72 years, 50 percent male). Children were tested with one of several standardized tests (e.g., SAT8, MAT7, CAT); normal curve equivalents were used to provide scores on the same metric for all tests for reading, vocabulary, spelling, language, mathematics, science, and social science. Test information was obtained from school records. For Attention/Executive Functions, correlations ranged from .11 to .15 (median = .11). Language

correlations ranged from .41 to .50 (median = .47), Sensorimotor correlations ranged from .00 to .06 (median = .04), Visuospatial domain correlations ranged from .20 to .37 (median = 27), and Memory and Learning correlations ranged from .30 to .42 (median = .37; see Korkman *et al.*, 1998, pp. 198, 207).

- A series of studies with small clinical and nonclinical school-age samples were reported in which NEPSY scores were correlated with other neuropsychological assessments. Included in these studies were assessments of 1) 18 children (mean age = 8.28 years, *SD* = 1.41 years, 61 percent male) who received both the NEPSY and the Benton Neuropsychological Tests (Benton, Hamsher, Varney, & Spreen, 1983) as part of a clinical evaluation, 2) a clinical sample of 17 children (mean age = 8.35 years, *SD* = 1.41 years, 65 percent male) who received both the NEPSY and the Multilingual Aphasia Examination-Second Edition (MAE; Benton & Hamsher, 1989) as part of a neuropsychological assessment, and 3) a nonclinical sample of 27 children (mean age = 7.37 years, *SD* = 2.06 years, 63 percent male) who received both the NEPSY and the Children's Memory Scale (CMS; Cohen, 1997). Korkman *et al.* (1998) report a number of associations with each of these three instruments, including correlations between subtests with a "…high degree of similarity in content and presentation" (Korkman *et al.*, p. 207), as well as correlations between NEPSY domain scores and Benton test scores, MAE subtest scores, and CMS indices (see Korkman *et al.*, pp. 198, 208 to 211).
    - *Attention/Executive Functions*. A correlation of .28 was reported between one NEPSY subtest and a similar Benton test. Correlations between Attention/ Executive Functions domain scores and Benton tests ranged in absolute value from .01 to .61 (median = .20); correlations with MAE subtests ranged from .09 to .47 (all positive; median = .26); and correlations with CMS Indices ranged in absolute value from .16 to .57 (median = .38).
    - *Language*. Three correlations, .44, .48, and .76, were reported between similar NEPSY and MAE subtests. Correlations between Language domain scores and Benton tests ranged in absolute value from .08 to .65, with a median correlation (when considered in terms of absolute value) of .24. Correlations with MAE subtests ranged from .13 to .70 (all positive), with a median of .37. Correlations with CMS Indices ranged from .01 to .55 (all positive), with a median of .45.
    - *Sensorimotor*. One subtest correlated -.16 with a similar subtest from the Benton, while a second subtest correlated .52 with an MAE subtest tapping a similar skill. The range of correlations between Sensorimotor domain scores and Benton tests was .06 to .48 (absolute values; median = .19); absolute values of correlations with MAE subtests ranged from .02 to .42 (median = .22); and absolute values of correlations with CMS Indices ranged from .09 to .41 (median = .21).
    - *Visuospatial*. Two subtests were similar to subtests from the Benton. Correlations were .35 and .77. Absolute values of correlations between Visuospatial domain scores and Benton tests ranged from .09 to .66 (median = .20); correlations with MAE subtests ranged in absolute value from .07 to .62 (median = .22); and correlations with CMS Indices ranged from .15 to .49 (all positive; median = .34).
    - *Memory and Learning*. One subtest was correlated -.03 with a similar subtest from the Benton, one was correlated .01 with a subtest from the MAE, and 3 were correlated .36, .45, .56, and .60 with 4 subtests from the CMS. Correlations

between domain scores and Benton tests ranged in absolute value from .03 to .76 (median = .10); absolute values of correlations with MAE subtests ranged from .14 to .68 (median = .29); and correlations with CMS Indices ranged from .21 to .74 (all positive; median = .48).

- Associations between NEPSY domain scores as well as Attention/Executive Functions subtest scores and scores on three attention tests were examined in a clinical sample of children diagnosed with ADHD. The three tests were 1) Conners' Continuous Performance Test (CCPT; Conners, 1995), 2) the Auditory Continuous Performance Test (ACPT; Keith, 1994), and 3) the Screening Test for Auditory Processing Disorders (SCAN; Keith, 1986). Not all tests were given to all children; 27 children (mean age = 8.22 years, *SD* = 1.22 years, 59 percent males) were tested with the CCPT, 13 children were tested with the ACPT (mean age = 9.77 years, *SD* = 1.83 years, 85 percent males), and 28 children (mean age = 8.10 years, *SD* = 1.34 years, 61 percent males) were tested with the SCAN (see Korkman *et al.*, 1998, pp. 198, 213).
    - Correlations of Attention/Executive Functions domain scores with the three scales from the CCPT ranged from -.06 to -.09 for the CCPT Hit Reaction and Attentiveness subtests, respectively. Correlations with the two ACPT scale scores were -.28 for Inattention Errors and -.27 for Impulsivity Errors (high scores on this measure indicate inattention and impulsivity problems). Correlations with the three scales from the SCAN were -.06, .23, and .35 for the Competing Words, Auditory Figure Ground, and Filtered Words subtests, respectively. Absolute values of correlations between Attention/Executive Functions subtest scores and scores on the CCPT ranged from .01 to .38 (median = .10). Absolute values of correlations between NEPSY subtest scores and ACPT scores ranged in absolute value from .04 to .76 (median = .21). Absolute values of correlations between NEPSY subtest scores and SCAN scores ranged from .01 to .43 (median = .13).
    - Correlations of Language domain scores were -.07, -.07, and -.24 with CCPT Risk Taking, Hit Reaction, and Attentiveness scores, respectively, .37 and -.44 with ACPT Inattention Errors and Impulsivity Errors subtest scores, and .07, .13, and -.16 with SCAN Competing Words, Auditory Figure Ground, and Filtered Words subtest scores.
    - Sensorimotor domain scores correlated -.28, -.30, and .36 with CCPT Risk Taking, Attentiveness, and Hit Reaction subtest scores, respectively, -.14 and -.21 with ACPT Impulsivity Errors and Inattention Errors subtest scores, and .01, .20, and .38 with SCAN Filtered Words, Competing Words, and Auditory Figure Ground subtest scores.
    - Visuospatial domain scores correlated .31, -.43, and -.47 with CCPT Hit Reaction, Attentiveness, and Risk Taking subtest scores, .14 and -.31 with ACPT Inattention Errors and Impulsivity Errors subtest scores, and -.02, .10, and .24 with SCAN Competing Words, Filtered Words, and Auditory Figure Ground subtest scores.
    - Correlations of Memory and Learning scores were -.09, .17, and -.27 with CCPT Risk Taking, Hit Reaction, and Attentiveness subtest scores, .02 and -.10 with ACPT Inattention Errors and Impulsivity Errors subtest scores; and -.12, -.13, and -.17 with SCAN Auditory Figure Ground, Competing Words, and Filtered Words subtest scores.

- Correlations were reported between NEPSY domain scores and selected scale and composite scores on the Devereux Scales of Mental Disorders (DSMD; Naglieri, LeBuffe, & Pfeiffer, 1994), including the Conduct and Attention/Delinquency scales and the Externalizing, Internalizing, and Total Scale composites. A sample of 10 children without clinical diagnoses and 13 children diagnosed with either ADHD or LD (mean age = 8.22 years, *SD* = 1.86 years, 78 percent male) was assessed (see Korkman *et al.*, 1998, pp. 198, 214).
    - Correlations of Attention/Executive Functions domain scores with DSMD scores were -.50 with the Conduct Scale, -.43 with Attention/Delinquency, -.48 with Externalizing, -.33 with Internalizing, and -.40 with the Total Scale Composite.
    - Correlations of Language domain scores with DSMD scores were -.15, -.24, -.18, .03, and -.10 with Conduct, Attention/Delinquency, Externalizing, Internalizing, and Total Scale, respectively.
    - Sensorimotor domain scores correlated -.39, -.27, -.35, -.27, and -.35 with DSMD Conduct, Attention/Delinquency, Externalizing, Internalizing, and Total Scale, respectively.
    - Correlations of Visuospatial domain scores with DSMD scores were -.18 with Conduct, -.17 with Attention/Delinquency, -.17 with Externalizing, .02 with Internalizing, and -.08 with the Total Scale Composite.
    - Correlations of Memory and Learning scores with DSMD Conduct, Attention/Delinquency, Externalizing, Internalizing, and Total Scale scores were -.14, -.15, -.13, .06, and -.05, respectively.

Discriminant Validity

Korkman *et al.* (1998) reported a series of studies where small samples of children with known neurological or developmental disabilities differed significantly on NEPSY subtest and core domain scores, compared with control groups of children matched on the basis of age, sex, race/ethnicity, and parent education. Criteria for inclusion for all studies included having an IQ of 80 or higher and no other psychiatric or neurological disorders that might present confounds for group comparisons. The significance of the difference between clinical and control group mean scores on domain and subtest scores, and the percentages of each group whose scores placed them one or two standard deviations from the standardized mean, were reported. When compared with matched control groups,

- A group of 51 children diagnosed with ADHD (mean age = 8.74 years, *SD* = 1.88 years, 69 percent male) had significantly lower mean scores on all 5 NEPSY domains. Despite a significant difference on Visuospatial domain scores, however, there were no significant group differences for any of the Visuospatial subtests (see Korkman *et al.*, 1998, pp. 216, 219).
- A group of 23 children diagnosed with both ADHD and LD (mean age = 9.45 years, *SD* = 1.85 years, 80 percent male) had significantly lower scores on all NEPSY domains except Visuospatial (see Korkman *et al.*, 1998, pp. 216, 220).
- A sample of 36 children diagnosed with specific learning disabilities related to reading (mean age = 9.58 years, *SD* = 1.48 years, 50 percent male) had significantly lower scores on two of the five domains—Language and Memory and Learning (see Korkman *et al.*, 1998, pp. 216, 222).

- A sample of 19 children diagnosed with disorders related to language processing abilities (mean age = 6.84 years, *SD* = 1.30 years, 37 percent male) had significantly lower scores on all NEPSY domains except Visuospatial (see Korkman *et al.*, 1998, pp. 216, 225).
- A sample of 20 children diagnosed with autism (mean age = 9.17 years, *SD* = 2.25 years, 83 percent male) had significantly lower scores on two of the five NEPSY domain scores – Attention/Executive Functions and Memory and Learning. Despite the lack of significance differences in other domain scores, however, autistic children had significantly lower scores on 5 of 6 Sensorimotor subtests and on 3 of 4 Visuospatial subtests (see Korkman *et al.*, 1998, pp. 216, 227).
- A group of 10 children (mean age = 9.20 years, *SD* = 1.48 years, 70 percent male) diagnosed with Fetal Alcohol Syndrome had significantly lower scores on all NEPSY domains except Sensorimotor (see Korkman *et al.*, 1998, pp. 216, 229).
- A group of 8 children with Traumatic Brain Injury (mean age = 8.38 years, *SD* = 2.07 years, 75 percent male) had significantly lower scores on 4 of 5 NEPSY domains; the Visuospatial domain difference was not significant (see Korkman *et al.*, 1998, pp. 216, 231).
- A sample of 32 children with hearing impairment (mean age = 9.28 years, *SD* = 1.53 years, 47 percent male) were significantly lower on three domains – Sensorimotor, Visuospatial, and Memory and Learning. Due to modifications in testing procedures for hearing impaired children, no domain scores were computed for Attention/Executive Functions or for Language; however; an examination of subtests within these domains indicated significant differences on only 1 of 5 Attention/Executive Functions subtests, but for 2 of 3 Language subtests (see Korkman *et al.*, 1998, pp. 216, 232).

Reliability/Validity Information from Other Studies

Stinnett, Oehler-Stinnett, Fuqua, and Palmer (2002) conducted secondary analyses of published data from the standardization sample (Korkman *et al.*, 1998); specifically, Stinnett *et al.* conducted a principal axis factor analysis using the published correlation matrix for 5- to 12-year-old children. These authors concluded that the NEPSY core domain structure is better defined with a single factor, rather than with the five domains described by Korkman *et al.*. They further found that most NEPSY subtests demonstrated adequate or better specificity, but concluded that this specificity does not necessarily indicate that the different subtests tap unique skills or abilities. Stinnett *et al.*' conclusions regarding the NEPSY were that caution should be taken by practitioners when interpreting test results, both at the domain and subtest levels, and that further research on the usefulness of the NEPSY is needed.

Comments

- Information provided in the manual on the reliability of the NEPSY subtests and standard scores generally indicates good internal consistency and strong correlations in test-retest reliability. It is noted, though, that reported stability statistics sometimes showed almost half the children being place in another score range on second assessment. Scant information was provided regarding interrater reliability. Only three subtests were examined, and the form of interrater reliability provided applies only to the ability of two raters to apply scoring rules to the same recorded child responses. In addition, adjusting interrater correlations for "rater leniency" may have inflated these correlations. It would be useful to have additional information as to whether two independent raters have

similarly high levels of agreement when conducting or observing the same assessment while it is taking place.

- Additional information on reliabilities within sex and within ethnicity/race would also be useful.
- A great deal of information was presented relevant to the validity of the NEPSY domains and subtests. With respect to construct validity as assessed by examining associations among domains, correlations ranging from .34 to .65 for 3- to 4-year-olds, and from .18 to .46 for 5- to 12-year-olds suggests that the five domains do each provide some unique information on the child's neuropsychological functioning. However, when examining correlations among subtests both within and across domains, it is clear that there is great variability in the magnitude of these correlations, but this variability does not appear to be greater across domains than within domains. As noted above, Stinnett *et al.* (2002) concluded that data from the standardization sample may not support a five-domain structure.
- Overall, there does appear to be some evidence of the validity of the NEPSY as a measure of general cognitive and perceptual skills and deficits in children. Evidence for this is found in results of comparisons of NEPSY results with results from other neuropsychological tests, comparisons between WPPSI-R and NEPSY scores, and in comparisons of NEPSY scores of groups of children with and without known neurological or developmental problems. However, associations with achievement as assessed through grades and standardized achievement test scores appear to be quite low for all domains except Language. In addition, not all associations with other measures were in the expected direction. One particular example of this was associations between BSID-II and NEPSY scores. Many of the reported studies have very small sample sizes, and larger studies will be important for the future development of the NEPSY.
- None of the information provided by Korkman *et al.* (1998) speaks to the usefulness of the Attention/Executive Functions domain, other domains, or subtests of the NEPSY for assessments of approaches to learning in young children. Although conceptually linked, it is likely that a substantial amount of additional basic research will need to be conducted in order to provide stronger linkages to children's early school adjustment.

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

None found.

V. Adaptations of Measure

None found.

References for Approaches to Learning Measures

Baker, A., Piotrkowski, C., & Brooks-Gunn, J. (1998). The effects of the Home Instruction Program for Preschool Youngsters (HIPPY) on children's school performance at the end of the program and one year later. *Early Childhood Research Quarterly, 13*, 571-588.

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

Benton, A. L., Hamsher, K., Varney, N. R., & Spreen, O. (1983). *Contributions to neuropsychological assessment*. New York: Oxford University Press.

Benton, A. L., & Hamsher, K. deS. (1989). *Multilingual Aphasia Examination – Second Edition*. Iowa City, IA: AJA Associates.

Brooks-Gunn, J., Berlin, L. J., Leventhal, T., & Fuligni, A. (2000). Depending in the kindness of strangers: Current national data initiatives and developmental research. *Child Development, 71*, 257-267.

Cohen, M. (1997). *Children's Memory Scale*. San Antonio, TX: The Psychological Corp.

Conners, C. K. (1995). *Conners' Continuous Performance Test (CPT)*. Toronto: Multi Health Systems.

Dunn, L., & Dunn, L. (1997). *Peabody Picture Vocabulary Test – Third Edition*. Circle Pines, MN: American Guidance Service.

Gresham, F. M., & Elliott, S. N. (1990). *Social Skills Rating System: Manual*. Circle Pines, MN: American Guidance Service.

Keith, R. (1986). *SCAN: A screening test for auditory processing disorders*. San Antonio TX: The Psychological Corp.

Keith, R. (1994). *The Auditory Continuous Performance Test.* San Antonio, TX: The Psychological Corp.

Kochanska, G., Murray, K. T., & Coy, K. C. (1997). Inhibitory control as a contributor to conscience in childhood: From toddler to early school age. *Child Development, 68*, 263-277.

Kochanska, G., Murray, K. T., Jacques, T. Y., Koenig, A. L., & Vendergeest, K. (1996). Inhibitory control in young children and its role in emerging internalization. *Child Development, 67*, 490-507.

Korkman, M., Kirk, U., & Kemp, S. (1998). *NEPSY: A Developmental Neuropsychological Assessment. Manual.* San Antonio, TX: The Psychological Corp.

McCabe, L. A., Hernandez, M., Lara, S. L., & Brooks-Gunn, J. (2000). Assessing preschoolers' self-regulation in homes and classrooms: Lessons from the field. *Behavioral Disorders, 26*, 53-69.

McCabe, L. A., Rebello-Britto, P., Hernandez, M. & Brooks-Gunn, J. (2004). Games children play: Observing young children's self-regulation across laboratory, home, and school settings. In R. DelCarmen-Wiggins & A. Carter (Eds.), *Handbook of infant, toddler, and preschool mental health assessment*. New York: Oxford University Press.

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

Maccoby, E. E., Dowley, E. M., Hagen, J. W., & Degerman, R. (1965). Activity level and intellectual functioning in normal preschool children. *Child Development, 36*, 761-770.

Meisels, S. J., & Atkins-Burnett, S. (1999). *Social Skills Rating System field trial analysis report and recommendations*. Final project report prepared for the National Opinion Research Center.

Naglieri, J. A., LeBuffe, P. A., & Pfeiffer, S. I. (1994). Devereux Scales of Mental Disorders. San Antonio, TX: The Psychological Corp.

Pittman, L. D., Li-Grining, C. P., & Chase-Lansdale, P. L. (2002). *Self regulation of economically disadvantaged children: The challenges and triumphs of measurement in the home.* Paper presented at New Directions in Young Children's Socio-Emotional Measures, a workshop co-sponsored by the NICHD Research Network on Child and Family Well-Being, NIMH, and the Science and Ecology of Early Development (SEED) initiative, Washington, DC, November 13.

Stinnett, T. A., Oehler-Stinnett, J., Fuqua, D. R., and Palmer, L. S. (2002). Examination of the underlying structure of the NEPSY: A developmental neuropsychological assessment. *Journal of Psychoeducational Assessment. 20*, 66-82.

Wechsler Individual Achievement Test. (1992). San Antonio, TX: The Psychological Corp.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intellingence—Revised*. San Antonio, TX: The Psychological Corp.

Wheeler, C., & Henrich, C. (2003). *Measuring the motivation of Head Start students: Validating the EZ-Yale Personality/Motivation Questionnaire.* Poster presented at the Biennial Meeting of the Society for Research in Child Development, Tampa, FL, April.

Zigler, E., Bennett-Gates, D., Hodapp, R., & Henrich, C. C. (2002). Assessing personality traits of individuals with mental retardation. *American Journal on Mental Retardation, 107*, 181-193.

General Cognitive Measures

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, pp. 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, pp. 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 Early Childcare Research Network, 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 Early Childcare Research Network, 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 Early Childcare Research Network, 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, *et al.*, 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.)[2]

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

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 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.3

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).

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

V. Adaptations 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).[4]

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

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 the author

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, pp. 76-77).

Another study was conducted with 36 3-, 4-, and 5-year-old children diagnosed with a cognitive deficit and a delay in at least one other area (communication, social, adaptive, behavior, or motor). 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, pp. 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 Early Childcare Research Network, 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 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 approaches (a work-first and an education-first approach) were evaluated in each of three sites (Atlanta, Georgia; Grand Rapids, Michigan; and Riverside, California) for a total of six JOBS programs evaluated. Children were between the ages of 5 years and 7 years at the two year follow-up. The 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 than did children in the control group. This study also examined the proportion of children in the two 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 program group children scored in the top quartile, compared to control group children. In addition, in one of the six programs, program group children were less likely to score in the bottom quartile on the SRC than were control group children.

V. Adaptations of Measure

Spanish Version

Description of Adaptation

A Spanish version of the BBCS-R is available. Spanish-language forms are designed for use with the English-language stimulus manual. The Spanish version is used as a curriculum-based measure only; 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.

Kaufman Assessment Battery for Children (K-ABC)

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.

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

As described by the authors

"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 Global Scales, preschool split-half reliabilities 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 results of principal components 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 K-ABC subtests were correlated with Successive and Simultaneous factor scores on the Das-Kirby-Jarman Successive-Simultaneous Battery (D-K-J; 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 p. 110).
    - The K-ABC Sequential Processing scale correlated .69 and .50 with the 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 *et al.*, 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.

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 the author

According to Markwardt (1998, p. 3), "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."

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

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, 1981) 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.

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.

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 the authors

"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; 1996) 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 & Levine, 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—the fall of 1988 and 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.

Types of questions asked vary by scale.
- *Spatial.* Children are asked to identify sequences using graphics (sequencing), and to identify pictures that represent two objects being combined into one (spatial integration).
- *Memory*. Children are asked to remember where things are in a circle (spatial memory). Other items involve the examiner showing and naming pictures, then asking children to find the named pictures in their test booklets (associative memory).
- *Concepts.* Children are asked to identify things that go together or that are the same (category concepts). Spatial concepts are examined by showing a picture in a box and then asking the child to identify another picture that goes best with the picture in the box.
- *Verbal.* Children are asked to identify the things that the examiner names (object naming), and are asked to identify which of several similar pictures matches a statement read by the examiner (syntax).

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; 1996). 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 (1992) and the CTBS/4 (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).5 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).

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

Stanford-Binet Intelligence Scale, Fourth Edition (SB-IV)[6]

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.

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

As described by the authors

"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.

[6] A fifth edition of the Stanford Binet (SB-V) was released in the spring of 2003, after the development of this profile.

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 include naming pictures, answering questions (e.g., "Why do people use umbrellas?"), and identifying what is absurd about a picture (e.g., a picture of a bicycle with square 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 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 data for ages 2 years to 5 years. At age 2, the lowest reliability found was .74 for the Copying subtest. All other reliability estimates were above .80; the highest estimate was .88 for the Memory for Sentences subtest. At age 3, reliabilities ranged from .81 for both Pattern Analysis and Copying to .91 for Absurdities. At age 4, reliabilities ranged from .81 for

Vocabulary to .88 for both Absurdities and Copying. At age 5, reliabilities ranged from .82 for Vocabulary to .90 for Pattern Analysis (see Thorndike, Hagen, & Sattler, 1986b, pp. 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, reliabilities were consistently high: .95 at age 2, .96 at age 3, and .97 at both ages 4 and 5 (see Thorndike, Hagen, & Sattler, 1986b, pp. 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, pp. 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 included 75 children 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 (see Thorndike, Hagen, & Sattler, 1986b, 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 included 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 both measures.
- 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) assessed a sample of at-risk preschoolers with both the SB-IV and the Battelle Developmental Inventory (BDI; Newborg, Stock, Wnek, 1984). The sample included 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 one standard deviation (SD) below the mean; in a second set of analyses, they used a two SD below the mean threshold. Saylor *et al.* 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

- Data for subtest, Area, and Composite scores indicate excellent internal consistency 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).
- Information provided in the Manual also provides support for the test-retest 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).
- 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 games and activities designed 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 activities that parents were taught to carry out with their children. Parent groups 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 ages 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.
- **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 Stanford-Binet Intelligence Scale than control group children. 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.

[7] See www.fpg.unc.edu/~ABC/embargoed/executive_summary.htm.
[8] See http://www.highscope.org/Research/PerryProject/perrymain.htm for a description of the study.

Wechsler Preschool and Primary Scale of Intelligence – Third Edition (WPPSI-III)

I. Background Information

Author/Source

Source: Wechsler, D. (2002a) *Wechsler Preschool and Primary Scale of Intelligence – Third Edition (WPPSI-III) Administration and Scoring Manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2002b) *Wechsler Preschool and Primary Scale of Intelligence – Third Edition (WPPSI-III) Technical and Interpretive Manual.* San Antonio, TX: The Psychological Corporation.

Publisher: The Psychological Corporation
19500 Bulverde Road
San Antonio, TX 78259
Website: www.psychcorp.com

Purpose of Measure

As described by the author
"As a psychoeducational tool, the WPPSI-III can be used to obtain a comprehensive assessment of general intellectual functioning. It can also be used as part of an assessment to identify intellectual giftedness, cognitive developmental delays, and mental retardation, and the results can serve as a guide for placement decision in clinical and/or school related programs" (Wechsler, 2002a, p. 6).

Population Measure Developed With

The WPPSI-III standardization sample included 1,700 children from across the United States. Selection was stratified based on 2000 U.S. Census data for age, sex, race, parent education level, and geographic location.

- Children were divided into nine age groups, ranging from 2 years, 6 months to 7 years, 3 months. Children between the ages of 2 years, 6 months and 6 years were grouped at 6-month intervals (e.g., 2 years, 6 months through 2 years, 11 months). The remaining two age groups were comprised of children between 6 years and 6 years, 11 months and children 7 years of age and older. There were 100 children in the oldest age group, 200 children in each of the other groups.
- Equal proportions of male and female participants were included in each age group.
- For each age group, the proportion of white, black, Hispanic, Asian, and other racial groups was based on the distribution reported by the 2000 U.S. Census.
- Five levels of parent education, based on the number of years of school attended, were used for stratification purposes. These levels were 0 to 8 years, 9 to 11 years, 12 years (i.e., high school graduation or equivalent), 13-15 years (i.e., partial college or associate's degree), and 16 or more years (i.e., bachelor's degree or higher). If the child resided with more than one parent or guardian, education levels were averaged.

- Geographic representation in the sample was broken into four sections of the United States (i.e., Northeast, South, Midwest, and West).
- Children were excluded from the standardization sample if 1) they had been tested by any intelligence measure in the previous 6 months, 2) they had uncorrected visual or hearing impairments, 3) they did not speak English, 4) they exhibited upper extremity motor impairments, 5) they were currently in hospital, mental or psychiatric wards or were taking medications that might depress test-performance, or 6) they had been previously diagnosed with a physical condition or ailment that could affect test-performance (e.g., stroke, epilepsy, brain tumor, etc.).

Age Range Intended For

- Ages 2 years, 6 months through 7 years, 3 months.

Key Constructs of Measure

There are three core constructs included in the WPPSI-III: Full Scale IQ, Verbal IQ, and Performance IQ. Each construct is based on a composite score derived from multiple subtests administered to the child. The subtests used to compute these three constructs are considered core subtests and vary based on child age. The WPPSI-III also offers supplemental subtests for varying age groups to provide more detailed information about child ability. Supplemental composite scores include the General Language Quotient for children between the ages of 2 years, 6 months and 3 years, 11 months, and the Processing Speed Quotient for children ages 4 years through 7 years, 3 months. The supplemental subtests can be used as substitutes for core subtests. Two optional subtests are offered for children ages 4 years and older; these subtests can be combined into a General Language Composite. Unlike Supplemental composite subtests, optional subtests cannot be substituted for core subtests.

- *Verbal IQ.* This scale is a composite of 4 core subtests that tap a range of verbal skills involving comprehension, reasoning, attention to verbally presented stimuli, and acquired knowledge. Verbal IQ subtests vary by child age.
    - *Information* (2 years, 6 months through 7 years, 3 months). Assesses children's ability to understand, remember, and recall factual information, and taps memory skills, crystallized intelligence, school knowledge, auditory comprehension, and verbal expression.
    - *Receptive Vocabulary* (2 years, 6 months through 3 years, 11 months). Assesses children's ability to comprehend verbal requests and respond to auditory stimuli, and taps auditory comprehension, and phonological memory. The supplemental subtest Picture Naming (see below) may be substituted for this subtest if deemed necessary. Receptive Vocabulary is also an optional subtest for children ages 4 years through 7 years, 3 months.
    - *Vocabulary* (4 years through 7 years, 3 months). Assesses children's knowledge and comprehension of words, and taps aspects of general knowledge, auditory perception, abstract thinking and verbal expression.
    - *Word Reasoning* (4 years through 7 years, 3 months). Assesses child's language, reasoning, verbal abstraction, and synthesis.
    - *Supplemental subtests*
        - *Picture Naming* (2 years, 6 months through 3 years, 11 months). Assesses children's ability to express language and associate visual stimuli with

language. It can be used as a substitute for Receptive Vocabulary. Picture Naming is also an optional subtest for children ages 4 years through 7 years, 3 months.

- *Comprehension* (4 years through 7 years, 3 months). Assesses children's verbal reasoning and ability to demonstrate practical information, and taps common sense and knowledge of "standards of behavior" (Wechsler 2002b, p. 26). This subtest may be substituted for one core Verbal subtest.
- *Similarities* (4 years through 7 years, 3 months). This supplemental subtest may be substituted for one core Verbal subtest. It taps children's verbal reasoning and concept formation.

- *Performance IQ.* This measure is a composite of 4 core subtests that tap a range of nonverbal problem solving, perceptual organization, and visual-motor proficiency skills. Performance IQ is a measure of "fluid reasoning, spatial processing, attentiveness to detail, and visual-motor integration" (Wechsler 2002b, p.136). Subtests within this construct vary by child age.
  - *Block Design* (2 years, 6 months through 7 years, 3 months). Assesses children's ability to analyze abstract visual stimuli, and taps non-verbal concept formation, hand-eye coordination, perception, and learning.
  - *Object Assembly* (2 years, 6 months through 3 years, 11 months). Assesses children's ability to see part-whole relationships, nonverbal reasoning, and trial and error learning.
  - *Matrix Reasoning* (4 years through 7 years, 3 months). Assesses children's fluid intelligence and general nonverbal intellectual ability.
  - *Picture Concepts* (4 years through 7 years, 3 months). Assesses children's abstract categorical reasoning.
  - *Supplemental Subtests*
    - *Object Assembly* (4 years through 7 years, 3 months). Assesses children's ability to see part-whole relationships, nonverbal reasoning, and trial and error learning. This supplemental subtest may be substituted for one core Performance subtest.
    - *Picture Completion* (4 year through 7 years, 3 months). Assesses children's visual-perceptual organization, ability to recognize object details visually, and concentration. This supplemental subtest may be substituted for one core Performance subtest.
- *Full Scale IQ: Full* Scale IQ is a composite score based on combined Verbal IQ and Performance IQ scores. Full Scale IQ is the most reliable measure of general intellectual functioning (*g*).
- Supplemental Scales
  - *General Language Quotient* (2 year, 6 months – 3 years, 11 months). A composite of the *Picture Naming* and *Receptive Vocabulary s*ubtests. It measures both receptive and expressive language. The General Language Quotient is also an optional subtest composite for children ages 4 years, through 7 years, 3 months.
  - *Processing Speed* (4 years through 7 years, 3 months). A composite of the Coding and Symbol Search subtests. It "provides a measure of the child's ability

to quickly and correctly scan, sequence, or discriminate simple visual information" (Wechsler 2002b, p. 136).

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

Comments
The standardization sample for the WPPSI-III appears to have been carefully selected so as to be nationally representative. Parental education, race/ethnicity and representation of geographic region all closely approximate characteristics of the U. S. population according to 2000 Census data.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
Child tasks vary by age and subtest.

- Verbal IQ tasks require children to point to illustrations of objects that are named or described, answer open–ended questions that might require knowledge of definitions, translate outside knowledge into language, express reasoning skills orally, and articulate more complex knowledge (e.g., understanding social conventions).
- Performance IQ tasks require children to use a set of colored blocks to match a block design presented by the test administrator, pick illustrations that are conceptually similar, put together puzzle pieces that resemble an illustration when joined, and look at a matrix of similar items and find which of 4 or 5 possible responses matches with the rest.

Who Administers Measure/Training Required?
Test Administration
Test administrators must have training and experience in the use of individual, standardized, clinical instruments. Such training should include experience with children whose characteristics (e.g. age, ethnic and racial background, language background, family socioeconomic characteristics, educational experience) are similar to the characteristics of the children being tested. Under some circumstances, trained teachers or examiners with supervision can administer the WPPSI-III.

Data Interpretation
Results should only be interpreted by individuals with the appropriate graduate or professional training and experience in psychological assessment.

Setting (e.g., one-on-one, group, etc.)
The WPPSI-III is designed to be administered in a one-on-one setting.

Time Needed and Cost
Time
Administration time will vary based on the subtests used, child's age, ability, motivation, and attention (see Wechsler, 2002a, p. 17).

- Fifty percent of children ages 2 years, 6 months through 3 years, 11 months complete the core subtests within 29 minutes, and 90 percent complete these subtests within 45 minutes.
- Fifty percent of children ages 4 years through 7 years, 6 months complete the core subtests within 41 minutes, and 90% complete these subtests within 61 minutes.

Cost

- All materials (w/out carrying case): $725
- Technical and Interpretative Manual: $48
- Administrative Manual $100

III. Functioning of Measure

Reliability Information from the Manual
Internal consistency
Split-half reliability was analyzed for all subtests with the exceptions of Coding and Symbol Search. These were excluded because they are speeded tests. All items were first rank ordered based on Item Response Theory difficulty estimates (i.e., rank order of items based on the actual performance of the sample). Odd and even items were then split to create two half-tests. Scores for the two half-tests were then correlated. These correlations were adjusted using the Spearman-Brown formula to derive internal consistency estimates for the full subtests. Across ages, average reliabilities for these subtests ranged from .84 (Block Design) to .95 (Similarities, administered to children ages 4 years and older). These reliabilities did not appear to vary systematically by child age. Across all ages, reliabilities for Verbal, Performance, and Full Scale IQs were .95, .93, and .96, respectively. Internal consistency of the General Language supplemental scale (an optional scale for children ages 4 years and older) was .93 (Wechsler, 2002b, p. 53).

Internal consistency reliability was also analyzed for various special populations, including 1) children with mental retardation, 2) children with developmental delays, 3) children with developmental risk factors, 4) autistic children, 5) children with Expressive Language Disorder, 6) children with Mixed Receptive-Expressive Language Disorder, 7) children with Limited English Proficiency, 8) children diagnosed with Attention Deficit/Hyperactivity Disorder, 9) children with motor impairment, and 10) intellectually gifted children. Sample sizes varied by population, ranging from as many as 70 children in the gifted sample to as few as 16 children in the motor impaired sample. Within the specialized populations there were also slight sample size differences between subtests. This was most pronounced for those with developmental

delays and developmental risk factors (see validity section below for full sample descriptions). Overall, average internal consistencies for subtests were comparable to or higher than those in the standardization sample (see Wechsler, 2002b, p. 55).

- Internal consistency reliabilities for children with mental retardation ranged from .89 (Comprehension) to .96 (Receptive Vocabulary).
- Reliabilities for children with developmental delays ranged from .89 (Comprehension) to .97 (Information).
- Reliabilities for children with developmental risk factors ranged from .75 (Vocabulary) to .98 (Picture Naming).
- Reliabilities for autistic children ranged from .94 (Block Design) to .98 (Similarities).
- Reliabilities for children with Expressive Language Disorder ranged from .73 (Information) to .98 (Similarities).
- Reliabilities for children with Mixed Receptive-Expressive Language Disorder ranged from .71 (Block Design) to .99 (Similarities).
- Reliabilities for children with Limited English Proficiency ranged from .82 (Picture Naming) to .96 (Matrix Reasoning and Word Reasoning).
- Reliabilities for children diagnosed with Attention Deficit/Hyperactivity Disorder ranged from .79 (Block Design) to .97 (Word Reasoning).
- Reliabilities for children with motor impairment ranged from .72 (Matrix Reasoning) to .98 (Object Assembly).
- Reliabilities for intellectually gifted children ranged from .61 (Object Assembly) to .91 (Picture Concepts). Internal consistencies were generally lower in this group, particularly for supplemental and optional subtests. The author suggested that lower reliabilities might result from "…restriction of range in the intellectually gifted sample and relatively lower ceilings for supplemental and optional subtests" (Wechsler, 2002b, p. 54).

Test-retest reliability

Test-retest reliability was assessed with a subsample of 157 children from the standardization sample (between 13 and 27 from each of the 9 age groups) who were assessed with the WPPSI-III on two occasions. Intervals between assessments ranged from 14 to 50 days, with an average interval of 26 days. This subsample was 39.5 percent female and 60.5 percent male, 66.2 percent white, 12.1 percent black, 17.2 percent Hispanic, and 4.5 percent another race. Parental education varied from 8.2 percent with between 9 and 11 years of schooling, 26.8 percent with 12 years, 53.5 percent with 13 to 15 years, and 11.5 percent with 16 years of schooling or more. The subsample was broken into three age bands: 2 years, 6 months to 3 years, 11 months, 4 years to 5 years, 5 months; and 5 years, 6 months to 7 years, 3 months. Test-retest correlations, corrected for variability in the standardization sample, were reported within age groups as well as for the full subsample.

- For WPPSI-III subtests, corrected test-retest correlations for youngest age group ranged from .74 for Object Assembly to .92 for Picture Naming (a supplemental subtest for this age group). Scale test-retest correlations were .84 for Performance IQ, .90 for Verbal IQ, and .92 for both Full Scale IQ and the supplemental General Language Scale (see Wechsler, 2002b, p. 60).
- For the middle age group (4 years to 5 years, 5 months), corrected test-retest correlations for subtests ranged from .69 for Block Design to .93 for Similarities (a supplemental subtest for this age group); all other subtest correlations ranged from .75 to .88. Test-

retest correlations for scales were .87 for Performance IQ, .90 for the optional General Language Scale, .92 for both Verbal and Full Scale IQ, and .93 for the supplemental Processing Speed Quotient (see Wechsler, 2002b, p. 60).

- For the oldest age group, corrected test-retest correlations for subtests ranged from .72 for Object Assembly (a supplemental subtest for this age group) to .86 for Information and Picture Naming (an optional subtest for this age group). Test-retest correlations for scales were .85 for the supplemental Processing Speed Quotient, .87 for Performance IQ, .91 for Verbal IQ, and .92 for both Full Scale IQ and the optional General Language Scale (see Wechsler, 2002b, p. 61).
- For all ages combined, corrected test-retest correlations for subtests ranged from .74 for Object Assembly to .90 for Similarities. Scale test-retest correlations were .86 for Performance IQ, .90 for the Processing Speed Quotient (ages 4 years and over), .91 for Verbal IQ and the supplemental/optional General Language Scale, and .92 for Full Scale IQ (see Wechsler, 2002b, p. 61).

Interscorer Agreement

All protocols for the norming sample were scored independently by two different scorers, in order to assess interscorer agreement. According to the author, "…the scoring criteria for most of the subtests are simple and objective, [and] interscorer agreement is very high, ranging from .98 to .99" (Wechsler, 2002b, p. 62). Special studies were conducted for three subtests that were identified by the authors as requiring more subjective interpretation: Vocabulary, Similarities, and Comprehension. For these three subtests, interscorer agreement was assessed with a subsample of 60 randomly selected cases from the norming sample. Four Master's level graduate students in School Psychology were trained to score WPPSI-III protocols; they subsequently each scored the 60 protocols independently. Reliabilities (intraclass correlations) were. 92, .90, and .95 at the item level for the Vocabulary, Similarities, and Comprehension subtests, respectively, and .97, .99, and .97 at the subtest score level (pp. 62-63).

Validity Information from the Manual

Wechsler (2002b) presents validity information for the WPPSI-III in a different way than in most other manuals, following recommendations found in the Standards for Educational and Psychological Testing developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA, APA, & NCME, 1999).[9] These Standards recommend that validity be seen as a unitary construct; there are different lines of evidence for the validity of a measure, rather than different, discrete types of validity. "Validity refers to the degree to which evidence supports the interpretation of test scores for their intended purposes" (Wechsler, 2002b, p. 69).

Evidence Based on Test Content

Wechsler (2002b, pp. 9-30, 70) indicates that items and subtests of the WPPSI-III were designed to tap "…a broad range of cognitive domains, including verbal reasoning, concept formation, sequential processing, auditory comprehension, cognitive flexibility, social judgment, perceptual organization, and psychomotor processing speed…" (p. 70). In the course of revising the WPPSI-R to create the WPPSI-III, a comprehensive literature review and consultations with

[9] Wechsler cites this as AERA (1999).

experts were conducted in order to ensure that items and subtests adequately covered these domains of cognitive functioning.

WPPSI-III development also involved a comprehensive system of obtaining expert opinion and feedback. An advisory committee was constructed from a diverse array of test administration and theory experts for this purpose. The committee was then consulted in regard to WPPSI-III goals, and ways in which the WPPSI-R could be improved. The advisory committee was included in every stage of development: concept development, piloting, national tryout, and finally standardization. The committee made recommendations at each stage. Outside the work of the advisory committee, focus groups and telephone surveys were done with practitioners using the WPPSI-R to address ways in which the former version could be improved. Surveys were also given to experts and examiners involved in the pilot, national tryout, and standardization stages of development to address possible improvements (see Wechsler, 2002b, p. 32).

Evidence Based on Response Processes

An additional source of evidence for the validity of a test involves determining whether the items and subtests measure the skills and abilities that they are intended to measure, i.e., that the child is using the expected response processes and not other unintended processes when responding to tasks. Much of the evidence for this, according to Wechsler (2002b, pp. 70-71) is based on the extensive literature review and consultation with experts conducted during the development of the WPPSI-III. Child responses to WPPSI-III items were also examined during measurement development for items that were frequently answered incorrectly. These items were examined to determine whether particular incorrect responses were plausibly correct. In early stages of item development, children were also asked to explain the reasoning behind their responses. When items appeared to be consistently missed and justification for the incorrect answer did not fit the construct's intent, they were modified or dropped from the test.

Evidence Based on Internal Structure

Several approaches were taken to provide evidence of validity based on internal structure, including intercorrelation studies, exploratory factor-analysis, and confirmatory factor analysis.

- *Intercorrelation studies*. A number of hypotheses were made regarding correlations among WPPSI-III subtests. Among these were that all subtests of the WPPSI-III should be correlated to some extent with each other because they are all thought to reflect elements of a general intelligence factor (*g*). At the same time, however, subtests contributing to the same scale (e.g., Verbal, Performance) were expected to be more highly correlated with each other than with scores for subtests contributing to other scales.
    - For ages 2 years, 6 months through 3 years, 11 months, all subtests were found to have statistically significant correlations with each other. These correlations ranged from .36 to .74 (see Wechsler, 2002b, p. 74).
        - The two Verbal subtests, Information and Receptive Vocabulary had a higher correlation with each other (.71) than with the two Performance subtests, Block Design and Object Assembly (correlations ranged from .36 to .44).

        - Block Design and Object Assembly were correlated .41 with each other, however, which was not consistently higher than correlations between these two subtest and the two Verbal subtests.
    - All subtests were significantly correlated with each other in the older age group as well (ages 4 years and older); correlations ranged from .27 to .74 (see Wechsler, 2002b, p. 75).
        - Correlations among the three Verbal subtests, Information, Vocabulary, and Word Reasoning, ranged from .69 to .74, consistently higher than correlations between the Verbal subtests and the three Performance subtests, Block Design, Matrix Reasoning, and Picture Concepts (correlations ranged from .44 to .51) or the two Processing Speed subtests, Coding and Symbol Search (correlations ranged from .29 to .42).
        - The two Processing Speed subtests were correlated .59, which was consistently higher than correlations with Verbal subtests as well as Performance Subtests (correlations ranged from .32 to .50).
        - As with the younger age group, however, Performance Subtests were not consistently more highly correlated with each other than with Verbal or Processing Speed subtests; correlations ranged from .41 to .51.
- *Exploratory Factor Analysis.* Multiple factor-analytic studies were done to assess the validity of the WPPSI-III constructs. One set of analyses included only core subtests, while a second set included both core and supplemental subtests. The current review discusses only analyses including supplemental subtests. The WPPSI-III technical and interpretive manual (Wechsler, 2002b, pp. 77-82) should be consulted for more detailed information.
    - A two-factor solution was specified for the youngest group (between the ages of 2 years, 6 months and 3 years, 11 months). Subtests loaded in the proposed way, with Receptive Vocabulary, Information, and Picture Naming loading together strongly (factor loadings ranging from .78 to .83) as the Verbal factor, and the Block Design and Object Assembly subtests loading together to form a Performance factor (factor loadings .59 and .56, respectively).
    - Analyses were conducted separately for children from 4 years through 7 years, 3 months. In addition, more restricted age bands were included in further analyses: between the ages of 4 and 4 years, 11 months; between the ages of 5 and 5 years, 11 months; and between the ages of 6 and 7 years, 3 months. A three-factor model was specified for these older groups (i.e., Verbal, Performance, Processing Speed; Wechsler, 2002b, pp. 79-82). For the full group of children ages 4 and older, subtests generally loaded in the proposed way. The Information, Vocabulary, Word Reasoning, Comprehension, and Similarities subtests loaded highly onto a Verbal factor (factor loadings ranging from .76 to .89), and the Symbol Search and Coding subtests loaded highly on a Processing Speed factor (factor loadings of .77 and .70, respectively). Block Design, Picture Completion, and Object Assembly subtests loaded as expected on the Performance construct (factor loadings of .62, .55, and .78, respectively). Matrix Reasoning also loaded most highly on the Performance factor, although the loading was a relatively low .36. The Picture Concepts subtest had relatively low loadings on all three factors, and had a higher loading on the Verbal factor (.30) than on the Performance factor

(.26). Results with the smaller age groups of older children were generally consistent with these findings for the full group of children ages 4 and older, with Matrix Reasoning and Picture Concepts subtests demonstrating relatively low loadings on the Performance factor and in some cases higher loadings on either or both the Verbal and Processing Speed factors.

- *Confirmatory Factor Analysis.* Confirmatory factor analyses were done to assess the relative goodness of fit of various models. Models included one- and two-factor solutions for children between 2 years, 6 months and 3 years, 11 months of age, and one-, two- and three-factor solutions for children ages 4 years and older. Multiple 3-factor models were tested allowing the Picture Concepts subtest to load onto the Verbal factor instead of or in addition to the Performance factor. For the youngest age group, goodness-of-fit was highest for the two-factor solution (i.e., Verbal and Performance). For children ages 4 years and older, the three-factor model in which Picture Concepts subtest scores were restricted to load only on the Performance factor and the three-factor model in which Picture Concepts scores were allowed to cross-load onto both the Verbal and Performance factors demonstrated the best fits to the data, with approximately equal goodness-of-fit statistics (Wechsler, 2002b, p. 86).

Evidence Based on Relationships with Other Variables

A series of analyses was conducted comparing children's scores on the WPPSI-III with scores on other measures, including the earlier version of the WPPSI, the WPPSI-R. The author provided detailed information regarding the demographic characteristics of each of these samples (see Wechsler, 2002b, p. 90 for these details).

- Correlations with the WPPSI-R. A sample of 176 children (48 percent female) between the ages of 3 years and 7 years, 3 months was assessed with both the WPPSI-R and the WPPSI-III in counterbalanced order, with a mean of 28 days between assessments (ranging from 8 to 58 days). A comparison of scale scores on the WPPSI-R and the WPPSI-III indicated that scores were consistently higher on the WPPSI-R; the difference between Performance IQ scores for the two tests was significant. Across the two tests, correlations between Verbal IQ, Performance IQ, and Full Scale IQ scores were .83, .68, and .82, respectively. Correlations between subtests appearing on both measures ranged from .51 to .76, with Verbal IQ subtests showing the strongest correlations between the two measures (Wechsler, 2002b, p. 91).
- Correlations with the WISC-III. Children's performance on the WPPSI-III and on the Wechsler Intelligence Scale for Children (WISC-III; Wechsler, 1991) was compared in a sample of 96 children (ages 6 years to 7 years, 3 months). The children were administered both tests (counterbalanced) an average of 24 days apart (with a range of 8 to 49 days). Mean scores were significantly higher for the WISC-III scale scores; the largest difference was 7.9 points between Performance IQ scores on the two measures. Correlations between the two measures' Verbal IQ, Performance IQ, and Full Scale IQ scores were .78, .74, and .85, respectively. Correlations between comparable subtests ranged from .45 to .65 (Wechsler, 2002b, p. 93).
- Correlations with the BSID-II. WPPSI-III scores were compared to Mental and Motor scale scores from the Bayley Scales of Infant Development, second edition (BSID-II; Bayley, 1993), in a sample of 84 children ages 2 years, 6 months to 3 years, 6 months. The children were administered each test, in counterbalanced order, with an average of

14 days between administrations (ranging from 7 to 42 days). There were no statistically significant differences between mean standard scores on the two measures (see Wechsler, 2002b, p. 95).

- Correlations between BSID-II Mental Scale composite scores and WPPSI-III scores ranged from .48 (Object Assembly) to .78 (Information) for subtests, and were .73, .61, and .80 with Verbal IQ, Performance IQ, and Full Scale IQ scores, respectively.
- Correlations between BSID-II Motor Scale scores and WPPSI-III scores were, as expected, consistently lower that the Mental Scale correlations. Correlations with WPPSI-III subtests ranged from .25 (Receptive Vocabulary) to .56 (Block Design). Correlations with Verbal IQ, Performance IQ, and Full Scale IQ scores were .33, .48, and .47, respectively.

- *Correlations with the DAS*. A sample of 164 children between the ages of 2 years, 3 months and 7 years, 3 months was assessed with both the Differential Abilities Scales (DAS; Elliott, 1990) and the WPPSI-III in counterbalanced order. There was a mean between-test interval of 28 days (ranging from 0 to 59 days).
  - Comparing standard scores on composites reflecting similar constructs from the two measures, scores were consistently higher on the WPPSI-III; however only one difference was significant. This difference of 3.8 points was between DAS Nonverbal Reasoning scores and WPPSI-III Performance IQ scores.
  - Correlations between the DAS Verbal, Nonverbal Reasoning, and General Conceptual Ability composite scores and the corresponding WPPSI-III Verbal IQ, Performance IQ, and Full Scale IQ scores were .78, .76, and .87, respectively (e.g., Wechsler, 2002b, p. 97).
  - The author also specifically highlights the patterns of correlations between subtests that are new to the WPPSI-III (i.e. Matrix Reasoning, Word Reasoning, Picture Concepts, and Picture Naming) and subtest and composite scores from the DAS as evidence of the construct validity of these new measures (see pp. 97-99). As would be expected, WPPSI-III Matrix Reasoning and Picture Concepts correlated .57 and .71 with Nonverbal Reasoning, respectively, while correlations were lower with Verbal composite scores (.45 and .52, respectively). In contrast, Word Reasoning and Picture Naming both correlated more highly with Verbal composite scores (correlations of .76 and .70, respectively) than with Nonverbal Reasoning scores (correlations of .47 and .37, respectively).
- Children's performance on the WPPSI-III was also compared to Wechsler Individual Achievement Test-II (WIAT-II, 2001) performance in a sample of 208 children between the ages of 4 years and 7 years, 3 months who completed both assessments. For most children, the WPPSI-III was administered first. There was an average of 14 days between administrations (ranging from 0 to 81 days).
  - Correlations between WPPSI-III Verbal IQ scores and WIAT-II composite scores ranged from .56 for Math to .77 for Total Achievement.
  - Correlations for Performance IQ scores with WIAT-II composite scores ranged from .36 with Written Language to .60 for Math.
  - Correlations between WPPSI-III Full Scale IQ scores and WIAT-II composite scores ranged from .62 with Written Language and .78 with Total Achievement (see Wechsler, 2002b, p. 100).

- Finally, children's scores on the WPPSI-III were compared to scores on the Children's Memory Scale (CMS; Cohen1997) in a sample of 40 children between the ages of 5 years and 7 years, 3 months. The mean between-test interval was 10 days (ranging from 0 to 38 days). The lowest correlations between WPPSI-III Verbal IQ, Performance IQ, and Full Scale IQ scores and CMS composites were consistently with Visual Delayed Index scores (correlations of .09, .14, and .10, respectively), while the highest correlations were consistently with CMS Attention/Concentration scores (correlations of .73, .68, and .79, respectively; Wechsler, 2002b, p. 102). Wechsler (p. 101) interprets these findings as indicating that "…Attention/Concentration (of all the CMS indices) is most closely related to *g* at these ages."

Special Population Studies

As noted in the reliability section of this profile, validity and reliability analyses were also done for various special population groups. In each study, the special population sample was compared to a matched control sample to determine whether expected group differences were evident.

- *Intellectually gifted children*. This sample included 70 intellectually gifted children (i.e., children whose scores on some other test of general cognitive abilities were at least 2 *SDs* above the mean) between the ages of 4 years, 6 months and 7 years, 3 months. As expected, the gifted children consistently scored higher on both composite scales and individual subtests of the WPPSI-III. Mean scores on Verbal, Performance, and Full Scale IQ were 125.8, 123.1, and 126.2, respectively, compared with scores of 107.4, 108.2, and 108.7 for the control group. Seventy-six percent of children in the gifted group earned Full Scale IQ scores of 120 or higher, whereas only 14 percent of the comparison group had scores at or above 120 (Wechsler, 2002b, pp. 105-106).
- *Children with mental retardation*. This sample included 59 children between the ages of 2 years, 6 months and 7 years, 3 months with mild to moderate mental retardation (i.e., children whose performance on a previously administered test of cognitive abilities was 2 to 4 *SDs* below the mean). Separate control groups were included for children with mild and moderate mental retardation. Children with either mental retardation generally performed more poorly on the WPPSI-III than did children in the control groups, and children diagnosed with moderate mental retardation performed more poorly than did those diagnosed with mild mental retardation. Mean Verbal, Performance, and Full Scale IQ scores were 58.1, 57.1, and 53.1, respectively, for children with moderate mental retardation, and 65.7, 65.6, and 62.1 for children with mild mental retardation. Means for the control groups on these three composites ranged from 90.7 (Performance IQ for the control group matched with the moderate mental retardation group) to 98.2 (Full Scale IQ for the group matched with the mild mental retardation group). Moreover, there was less individual variability in the mental retardation groups, compared with the control groups, and both mental retardation groups scored consistently lower than their control groups on all subtests. Almost 94 percent of children diagnosed with mild mental retardation had scores of 75 or lower, compared to 8 percent of the control group. Similarly, 75 percent of children diagnosed with moderate mental retardation had scores of 60 or less, while no children in the comparison group scored at this level (Wechsler, 2002b, pp. 107-112).

- *Children with developmental delays*. This sample included 62 children with cognitive or multiple developmental delays between the ages of 2 years, 3 months and 7 years, 3 months. The developmentally delayed children had significantly lower scores than did comparison group children on all composites. Verbal, Performance, and Full Scale IQ scores for the delayed group were 82.8, 86.1, and 81.8, respectively, while scores for the control group were 99.5, 101.2, and 99.9. Nineteen percent of children with developmental delays had Full Scale IQ scores of less than or equal to 70. Only one child in the matched group had such a low score (Wechsler, 2002b, pp. 113-114).
- *Children with developmental risk factors*. A group of 31 children ages 2 years, 6 months to 7 years, 3 months with developmental risk factors (e.g., low birth weight, brain hemorrhage, prenatal drug or alcohol exposure, a history of abuse or neglect) were included in these analyses. As expected, children with developmental risk factors had significantly lower scores on all WPPSI-III composite scores, with the exception of Processing Speed. Mean Verbal, Performance, and Full Scale IQ scores for the group with developmental delays were 88.6, 85.7, and 85.7, respectively, compared with 97.7, 96.5, and 96.4 for the control group (Wechsler, 2002b, pp. 112-114).
- *Children with autism*. Compared to a matched comparison group, a group of 21 autistic children between the ages of 3 years and 6 years, 11months had significantly lower scores on all WPPSI-III composites. Mean Verbal, Performance, and Full Scale IQ scores were 70.6, 88.2, and 76.6, respectively, compared with scores of 98.7, 99.5, and 98.6 for the control group. Further, as hypothesized by the author, autistic children's Verbal IQ scores were significantly lower than their Performance IQ scores (Wechsler, 2002b, pp. 116-117).
- *Children with Expressive Language Disorder (ELD)*. This group included 23 children with ELD between the ages of 3 years and 6 years, 11 month. Children with ELD had significantly lower Verbal IQ and Full Scale IQ scores (90.6 and 90.1, respectively) than did children in their comparison group (scores of 98.8 and 97.6, respectively). However, as expected, Performance IQ scores for the two groups did not differ significantly (scores of 92.9 and 96.9 for children with and without ELD, respectively (Wechsler, 2002b, p. 119).
- *Children with Mixed Receptive-Expressive Language Disorder (RELD)*. The WPPSI-III was administered to a group of 27 children ages 4 years through 7 years, 3 months diagnosed with RELD. In contrast to the findings for children diagnosed with Expressive Language Disorder, children in the RELD group had lower scores on all composite scales than did control group children. Mean scores for Verbal, Performance, and Full Scale IQ were 83.1, 85.2, and 81.9, respectively, for the RELD group, compared to 101.1, 95.6, and 96.9 for the comparison group (Wechsler, 2002b, p. 121).
- *Children with Limited English Proficiency (LEP)*. The WPPSI-III was administered to 44 LEP children between the ages of 3 years, 6 months and 7 years, 3 months who comprehended enough English to receive test instructions in English. As predicted, the LEP group performed more poorly than did control group children on all subtests and scales that measured verbal ability, but did not show any significant differences on Performance scales. Means for Verbal, Performance, and Full Scale IQ were, respectively, 80.2, 95.0, and 87.0 for the LEP group, compared with means of 96.1, 97.4, and 96.7 for the comparison group (Wechsler, 2002b, p. 123).

- *Children with Attention Deficit/Hyperactivity Disorder (ADHD)*. This sample included 41 children between the ages of 3 years, 6 months and 7 years, 3 months diagnosed with ADHD. The specific prediction was that the two groups would differ primarily on the Processing Speed Quotient. Contrary to this prediction, there were no significant differences between the groups. Score for Verbal IQ, Performance IQ, Full Scale IQ, and the Processing Speed Quotient were 93.8, 97.4, 94.3, and 95.4 respectively for the ADHD group, and 97.8, 101.2, 98.3, and 96.0 for the comparison group (Wechsler, 2002b, p. 125). In this sample, 21 of the ADHD children were currently medicated; Wechsler notes that "…investigations with separate samples of children with ADHD based on subtype are needed, as well as investigations comparing the performance of medicated and nonmedicated children with ADHD" (p. 124).
- *Children with motor impairments*. The WPPSI-III was administered to 16 children between 3 years and 6 years, 11 months of age with motor impairments. As hypothesized, children with motor impairments had significantly lower scores than did children in the control group on Performance IQ (mean scores of 87.7 and 108.1 for impaired and unimpaired children, respectively) and Full Scale IQ (mean scores of 94.2 and 106.4), but did not differ on Verbal IQ (mean scores of 102.2 and 105.0; Wechsler, 2002b, p. 127).

Reliability/Validity Information from Other Studies

Given the relative newness of the measure, we were unable to locate any other reliability or validity studies.

Comments

- Overall, information provided by Wechsler (2002b) indicates that the WPPSI-III subtests and composite scales demonstrate good reliability. In the standardization sample, reported split-half reliabilities were all high. Split-half reliabilities were also generally high for special population samples. Findings for intellectually gifted children do suggest that there may be problems with restrictions of range and consequent lowered reliabilities for some subtests with this population.
- Similarly, test-retest correlations were all high for all age groups, supporting the reliability of this measure.
- Interrater reliability was assessed only for the three tests that were identified by Wechsler as requiring subjective judgments on the part of the scorer. The reported intraclass correlations, all .90 or higher, indicated very high agreement between trained independent scorers.
- In the test-retest reliability analysis, all children were assessed twice by the same examiner. Interrater reliability analyses were designed to determine whether multiple scorers could apply standard scoring rules to the same protocols in order to arrive at the same scores. Thus, none of the reliability analyses provided in the manual can address the extent to which children would receive the same scores from two different evaluators conducting independent assessments.
- Overall, the validity evidence provided in the manual appears to be quite strong. Subtests appear to correlate as expected with each other, although this appears to be more strongly the case for Verbal IQ and Processing Speed subtests than for Performance IQ subtests. Further, WPPSI-III scores correlated in expected ways with scores on other measures of

cognitive abilities, including the WPPSI-R, the WISC-III, the BSID-II, the DAS, and the CMS.

- Some differences were noted between scores obtained using the WPPSI-III versus scores obtained with the former version of the measure, the WPPSI-R. These differences, when significant, showed generally moderate to small effect sizes, and correlations between the two measures were strong. However, findings of generally lower scores on the WPPSI-III than on the WPPSI-R suggest that the same version of the WPPSI should be used at all time points in studies attempting to assess change in individuals or groups.
- Additional support for the validity of the WPPSI-III comes from studies with special populations. In these studies, predicted group differences were generally found, with one exception: Processing Speed Quotients for a group of children diagnosed with ADHD did not differ from those of children in a non-ADHD control group. The author notes that half of these children were currently medicated, possibly explaining this lack of a significant difference. It is also worth noting that in the study comparing LEP with non-LEP children, there were no significant differences in Performance IQ, suggesting that this may be a valid measure for LEP children, even when the assessment is administered in English, so long as the children understand enough English to follow test instructions.

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

Given the relative newness of the measure, we were unable to locate any current empirical work using the WPPSI-III.

V. Adaptations of Measure

None.

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.

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 the authors

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 representative sample 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. For preschoolers and school-age children (K through twelfth grade), parents' education was also controlled.
- All participants 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.

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 form a Standard Battery, 10 others are included in an Extended Battery that 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.

- For the WJ III COG, examples of tasks include pointing to the picture of a word spoken by the examiner, identifying two or three pieces that form a complete target shape, listening to a series of syllables or phonemes and then blending them into a word, and pointing to the matching shapes in a row of four or five shapes.
- For the WJ III ACH, examples of tasks include identifying a printed letter, listening to and recalling details of a story, identifying an object, and solving 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).
    - 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, pp. 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, pp. 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, pp. 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, pp.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 the prior version of the measure, the WJ-R (Woodcock & Johnson, 1989), 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. Adaptations of Measure

Spanish Version of WJ III

A Spanish version of the WJ III is available.

References for General Cognitive Measures

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

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

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

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

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

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 Corp.

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*, 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.

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*, 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*, 339-357.

California Achievement Tests. (1992). Monterey, CA: CTB/McGraw Hill.

Cohen, M. (1997). *Children's Memory Scale*. San Antonio, TX: The Psychological Corp.

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

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 Corp.

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*, 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,* 237-248.

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

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*, 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*, 294-302.

Love, J. M., Kisker, E. E., Ross, C. M., Schochet, P. Z., Brooks-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*. 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 Corp.

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*, 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*, 1072-1077.

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

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*, 179-183.

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.

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. (1991). *The Wechsler Intelligence Scale for Children—Third Edition*. San Antonio, TX: The Psychological Corp.

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

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

Wechsler, D. (2002a). *Wechsler Preschool and Primary Scale of Intelligence – Third Edition (WPPSI-III) administration and scoring manual*. San Antonio, TX: The Psychological Corp.

Wechsler, D. (2002b). *Wechsler Preschool and Primary Scale of Intelligence – Third Edition (WPPSI-III) technical and interpretive manual*. San Antonio, TX: The Psychological Corp.

Wechsler Individual Achievement Test—Second Edition. (2001). San Antonio, TX: The Psychological Corp.

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*, 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 Corp.

Language Measures

Clinical Evaluation of Language Fundamentals – Preschool (CELF-Preschool)

I. Background Information

Author/Source

Source: Wiig, E., Secord, W., & Semel, E. (1992). *CELF- Preschool: The Clinical Evaluation of Language Fundamentals – Preschool Examiner's Manual*. San Antonio, TX: The Psychological Corporation.

Publisher: The Psychological Corporation
19500 Bulverde Road
San Antonio, TX 78259
Phone: 1-800-872-1726
Website: www.psychcorp.com

Purpose of Measure

As described by the authors
"Clinical Evaluation of Language Fundamentals – Preschool (CELF-Preschool) is a practical and efficient tool for identifying, diagnosing, and performing follow-up evaluations of language deficits in preschool children. CELF-Preschool, a downward extension of the Clinical Evaluation of Language Fundamentals- Revised (CELF-R), is an individually administered test that assesses receptive and expressive language ability. Like CELF-R, CELF-Preschool explores the foundations of language form and content: word meaning (semantics), word and sentence structure (morphology and syntax), and recall of spoken language (auditory memory)…The test can be used to provide valuable information to a variety of professionals involved in preschool education, including speech-language pathologists, child psychologists, educational diagnosticians, and special educators" (Wiig, Secord, & Semel, 1992, p. 1).

Population Measure Developed With

The development of the CELF-Preschool included various samples of children for different stages of test development. Data for an initial pilot study sample and national tryout sample were collected prior to the standardization sample. For the sake of brevity, sample characteristics will only be provided for the standardization sample. However, some further information about the field and tryout testing is provided in the construct validity section of this document.

- The standardization sample was nationally representative (based on 1988 update data from the 1980 Census). It consisted of 800 children from 42 states and was stratified by age, gender, race/ethnicity, parent education and geographic location.
- Fifty percent of the sample children were male, 69.6 percent were white, 14.9 percent were black, 11.6 percent were Hispanic, and 3.9 percent were considered "other." The ethnic breakdown of the sample differed only slightly (by a tenth of a percent) from the U.S. population in two instances (Hispanic, and other).
- When stratified by age, 100 children were selected for each of 8 age groups. Ages ranged between 3 years and 6 years, 11 months, with age groups consisting of children within 6-month spans (e.g., 3 years through 3 years, 5 months; 3 years, 6 months through 3 years, 11 months, etc.).

- Regarding education, 17.4 percent of the children's mothers had 11 years or less of education, 37.1 percent had 12 years of education, 26.4 percent had between 13 and 15 years of schooling, and 19.1 percent had 16 years or more. This distribution was comparable to the estimated distribution found in the United States for 1988.

Age Range Intended For

Ages 3 years through 6 years, 11 months.

Key Constructs of Measure

The CELF-Preschool contains three composite scales (Receptive Language, Expressive Language, and Total Language), with the Receptive and Expressive Language scales each being comprised of three subtests. The Total Language Scale is the total of the standard scores for all 6 subtests (i.e., the sum of the *Receptive* and *Expressive Language* scales). In addition, a Quick-Test can be administered as a screener to determine the need for further testing.

- *Receptive Language:* This scale contains the *Linguistic Concepts*, *Sentence Structure*, and *Basic Concepts* subtests.
    - The *Linguistic Concepts* subtest assesses understanding of concepts such as the use of conjunctions (e.g., and, or), positive versus negative (e.g., is, is not), and location in space or time (e.g., do X, *then* do Y; pick the bird *next to* the cow).
    - The *Sentence Structure* subtest taps understanding of early-acquired sentence formation rules, such as the ability to identify key attributes of items from an example of those items (e.g., an apple is food, pants are a piece of clothing).
    - The *Basic Concepts* scale involves the child's ability to understand modifiers, such as relative amount or size (e.g., more, less, big, small), as well as basic concepts such as same versus different, and inside and outside.
- *Expressive Language:* The Expressive Language scale is comprised of the *Recalling Sentences in Context*, *Formulating Labels*, and *Word Structure* subtests.
    - The *Recalling Sentences in Context* subtest measures the child's ability to recall and repeat a sentence that is read to him/her in the context of a story.
    - The *Formulating Labels* subtest focuses on the child's ability to give verbal labels to nouns and verbs depicted in illustrations.
    - The *Word Structure* subtest measures the child's understanding of morphological rules, through tapping his/her ability to provide word forms such as past tense, irregular verbs, and pronoun assignment.
- *Total Language: The* Total Language score is derived by summing scores for all six subtests included within the Expressive and Receptive Language scales.
- *Quick -Test:* The Quick-Test may be used as an initial step in assessment and consists of only the *Linguistic Concepts* and *Recalling Sentences in Context* subtests. A score of seven or below on the Quick -Test indicates that the remainder of the battery should be given to address specific language deficits.

Norming of Measure (Criterion or Norm Referenced)

Norm referenced.

Comments
It is unclear whether children with disabilities and/or primary languages other than English were included in the sample.

II. Administration of Measure

Who is the Respondent to the Measure?
- Child.

If Child is Respondent, What is Child Asked to Do?
- Child tasks vary depending upon the subtest being administered. Within the Linguistic Concepts subtest of the Receptive Language scale, children are asked to show their understanding of such concepts as the use of conjunctions, positive versus negative, and location in space or time, by pointing at illustrations that convey the meaning of the concept. The Sentence Structure subtest requires the child to point to an illustration that best matches a phrase the administrator says. The final subtest of this composite scale, Basic Concepts, uses a similar pointing task to that used in the Sentence Structure subtest. In this subtest, children are again provided with three illustrations to choose from and asked to identify the one that best illustrates a concept.
- Subtests of the Expressive Language scale also vary by the type task the child is asked to perform. The Recalling Sentences in Context subtest requires the child to recall specific sentences from a story and repeat them back verbatim, in response to a question from the administrator. Administrator questions take place directly after the probe sentence in the story is read. The Formulating Labels subtest taps the child's ability to provide the correct label for nouns and verbs in illustrations. Children are scored based on how closely their verbal responses resemble a target response. Lastly, the Word Structure subtest assesses the child's ability to describe what is being portrayed in an illustration.

Who Administers Measure/Training Required?
Test Administration
"Examiners should have had experience administering, scoring, and interpreting standardized tests before attempting to administer and interpret CELF-Preschool"(Wiig, *et al.*, 1992, p. 5).

Data Interpretation
Same as above.

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
Administration time depends on the age of the child, with younger children usually needing more time. Reported administration times ranged from 44 minutes for the youngest group to 30 minutes for the oldest children. The Quick-Test administration times are approximately half that of the full test.

Cost

- Stimulus Manual 1 (Linguistic Concepts, Formulating Labels, Sentence Structure, and Basic Concepts with easel.): $142.50.
- Stimulus Manual 2 (Recalling Sentences in Context): $37.00.
- Stimulus Manual 3 (Word Structure): $37.00.
- Record Forms, 25: $49.00.
- Examiner's Manual: $54.00.

Comments

The CELF-Preschool includes a Behavioral Observation Checklist to be used during or after the assessment to record specific child behaviors that occur in the testing session. The observation checklist makes note of physical activity level, attention to task, response latency, fatigue/boredom/frustration, and level of interaction. The Behavioral Observation Checklist may be used in conjunction with CELF-Preschool scores to gain a more complete picture of the child's performance. It is unclear whether there is a standardized way to include the Behavioral Observation Checklist in CELF-Preschool scoring.

III. Functioning of Measure

Reliability Information from the Manual

Internal consistency

Internal consistency reliability coefficients were calculated for each of the CELF-Preschool subtests and composite scales, based on data for the 800 children in the standardization sample (100 per age group). Reliability data were presented for children within six-month age bands, from the youngest (3 years through 3 years, 6 months) to the oldest (6 years, 6 months through 6 years, 11 months).

- Reliability coefficients for the Receptive Language scale ranged from .73 for the 6 years, 6 months through 6 years, 11 months age group, to .92 for the 4 years through 4 years, 5 months age groups. Coefficients for the Linguistic Concepts subtest ranged from .69 to .86 for the 6 years, 6 months through 6 years, 11 months and 4 years through 4 years, 5 months groups, respectively. Younger children tended to show slightly higher internal consistency reliabilities on this subtest. Basic Concepts coefficients ranged from .49 for children between the ages of 6 years and 6 years, 6 months, to .81 for the children between the ages of 4 years and 4 years, 5 months. The Sentence Structure subtest showed internal reliabilities between .30 for the oldest group of children and .83 for the group of children between the ages of 3 years, 6 months and 3 years, 11 months (Wiig, *et al.*, 1992, pp. 53-54).
- Internal consistency reliability coefficients for the Expressive Language scale ranged from .82 for the oldest group of children to .95 for the group of children between the ages of 4 years and 4 years, 5 months (as was seen for the Receptive Language scale). The Recalling Sentences in Context subtest showed a range in reliability coefficients between .72 for the oldest group of children (6 years, 6 months through 6 years, 11 months) and .93 for the two groups that comprised the age range of 3 years, 6 months through 4 years, 5 months. The same pattern held for the Formulating Labels subtest, with the oldest group of children showing the lowest reliability coefficient, .71, and the group of children

between the ages of 4 years and 4 years, 5 months showing a coefficient of .88. The Word Structure subtest also had the lowest coefficient for the oldest children (.64) and the largest coefficient for the children between the ages of 4 years and 4 years, 5 months (.88; Wiig, *et al.*, 1992, 1992, pp. 53-54).

- Reliability coefficients for the Total Language scale ranged from .86 for the oldest children, to .96 for children between the ages of 3 years, 6 months and 3 years, 11 months and the group of children between the ages of 4 years and 4 years, 5 months (Wiig, *et al.*, 1992, p. 54).

Test-Retest Reliability
Fifty-seven children (28 male, 29 female) participated in an examination of test-retest reliability. The sample was broken into two roughly equal age groups, 3 years, 6 months through 3 years, 11 months (n = 27) and 4 years, 6 months through 4 years, 11 months (n = 30). The CELF-Preschool was administered twice to each child by the same examiner. The time between assessment periods ranged from two to four weeks. For the younger group of children, correlations between the two administrations ranged between .60 for the Sentence Structure subtest to .92 for Formulating Labels. Composite scale test-retest correlations were generally higher than subtest correlations, ranging from .93 for Receptive Language to .97 for Total Language scores. Similar results were noted for the group of older children, with correlations ranging from .63 for Sentence Structure to .87 for the Formulating Labels subtest. Composite scales were higher, ranging from .93 for Total Language score to .97 for Receptive Language. The same patterns were noted as for the age subgroups when the two age groups were collapsed (Wiig, *et al.*, 1992, pp. 54-55).

Interrater Reliability
The manual notes that the Formulating Labels subtest is the only subtest that requires judgment on the part of the administrator. Because of this, interrater reliability was assessed for this subtest only. Three independent raters scored responses from a random sample of 75 children drawn from the standardization and validity studies. Mean interrater agreement across all items in the Formulating Labels subtest was 90 percent (Wiig, *et al.*, 1992, p. 55).

Validity Information from the Manual
Content Validity
The authors assert that the content validity of the CELF-Preschool is inherent in its measurement of morphology and syntax, which have been well documented in the research regarding "…language development, language disorders, and competent language use…" (Wiig, *et al.*, 1992, p. 55).

Construct Validity
Intercorrelations were calculated between CELF-Preschool Receptive and Expressive Language scales in each age group. These ranged between .60 and .84, with the strongest correlation found for children between the ages of 4 years and 4 years, 5 months, and the weakest for children between 6 years, 6 months and 6 years, 11 months of age. The authors interpret these findings to indicate that the two constructs are related but not entirely overlapping.

The authors also cite discriminant validity—the ability of the test to discriminate between children with established diagnoses of language disorders from children without such diagnoses—as an example of construct validity. The CELF-Preschool was administered to 80 children (54 male, 26 female) between the ages of 5 years and 6 years, 11 months who had been diagnosed with language disorders, and a matched sample of children without language disorders. CELF-Preschool scores of one standard deviation below the mean, and of one and a half standard deviations below the mean were compared as cutoffs for identifying children at elevated risk for language disorders. Using the cutoff of one standard deviation below the mean, the CELF-Preschool correctly identified children as having a language disorder or not having a language disorder 74 percent of the time. Using this cutoff, the CELF-Preschool was much more accurate in labeling children who did not have language disorders than in identifying those who did, missing 40 percent of the children who had a language disorder. The cutoff of one and a half standard deviations below the mean had a slightly worse overall accuracy of 71 percent. Using this cutoff, the CELF-Preschool classified 52.5 percent of children with language disorders as not having language disorders (Wiig, *et al.*, 1992, pp. 61-62).

Concurrent Validity

CELF-Preschool is a downward extension of the Clinical Evaluation of Language Fundamentals, Revised (CELF-R), designed for use with school-age children. Three of the CELF-Preschool subtests are nearly identical to those found in the CELF-R (i.e., Linguistic Concepts, Sentence Structure, Word Structure); the remaining three subtests are comparable but not identical (i.e., Basic Concepts, Recalling Sentences in Context, Formulating Labels). CELF-Preschool scores were compared to scores on the CELF-R in a sample of 80 language-disordered children between the ages of 5 years and 6 years, 11 months to measure comparability between the two versions of the assessment within this population. All children had been previously diagnosed as having a language disorder, with some exhibiting some evidence of other disorders as well (e.g., articulation, phonology, voice, or fluency problems). There were more males in the sample than females (54 and 26, respectively), and children were evenly divided into four age groups: 5 years to 5 years, 5 months; 5 years, 6 months to 5 years, 11 months; 6 years to 6 years, 5 months; and 6 years, 6 months to 6 years, 11 months. Each child was given the CELF-Preschool and the CELF-R, with order of administration counterbalanced.

- Correlations between the two assessments for the youngest group ranged from .41 for the Formulation Labels subtest to .84 for both the Linguistic Concepts subtest and the Receptive Language composite scale score.
- Correlations ranged from .41 for the Basic Concepts subtest and its CELF-R equivalent to .86 for the Total Language composite score in the two CELF versions for the group of children between the ages of 5 years, 6 months and 5 years, 11 months.
- A similar pattern was found for the next oldest group (between 6 years and 6 years, 5 months), with the weakest correlation (.31) between the Basic Concepts subtest and CELF-R equivalent, and the strongest (.93) between the Total Language composites scores on the two CELF versions.
- Correlations between the two tests ranged from .51 for Sentence Structure and .74 between the Receptive Language composite score for the oldest group of children.
- In each case, the lowest correlation was noted for a comparable, but not direct downward extension subtests. Thus, CELF-Preschool subtests that were a direct downward extension of the CELF-R were more highly correlated than subtests that were merely

> comparable between the two measures (i.e., Linguistic Concepts, Sentence Structure, and Word Structure are shared between the two). The authors note that the Expressive Language composite scores for the CELF-Preschool are significantly higher than those on the CELF-R, as might be expected given the greater difficulty of the latter and the diagnoses of child language disorder for the children in the sample (Wiig, *et al.*, 1992, pp. 56-57).

In another sample, 100 children between the ages of 5 years and 6 years, 11 months (50 male, 50 female) were assessed with both the CELF-Preschool and the CELF-R Screening Test, a criterion rated screener of language disorders. Instead of using standard scores for the CELF-Preschool, a criterion for "passing" was set at a Total Language score of 77 (1.5 standard deviations below the mean), and children were compared for how they were rated on each assessment. Overall there was 75 percent agreement in identifying children with and without language disorders using the screener and the CELF-Preschool. Most of the disagreement occurred when children were "missed" by the CELF-Preschool and "flagged" by CELF-R Screener (Wiig, *et al.*, 1992, p. 58).

In another study, the CELF-Preschool composite scales were compared to the Preschool Language Scales-Third Edition (PLS-3; Zimmerman, Steiner, & Pond, 1992) Auditory Comprehension, Expressive Communication, and Total Language scales in a sample of 53 children between the ages of 3 years and 4 years, 5 months. Both assessments were administered to each child (with order counterbalanced across children), with between-test intervals ranging from one day to two weeks. Correlations between the CELF-Preschool Receptive Language scale and the PLS-3 scales ranged from .73 to .83; the highest correlation was between the CELF-Preschool Receptive Language scale and both the comparable PLS-3 Auditory Comprehension scale and the Total Language scale. Correlations between the CELF-Preschool Expressive Language scale and PLS-3 scales ranged from .81 to.86, with the strongest relationship found with the PLS-3 Total Language score. It is noted that comparable scales of the CELF-Preschool and the PLS-3 were generally the most strongly related, with the exception of the CELF-Preschool Expressive Language scale that was equally related to both PLS-3 Auditory Comprehension and Expressive Communication scores. As might be expected, the CELF-Preschool Total Language scale was most strongly correlated with the PLS-3 Total Language score (correlation of .90; Wiig, *et al.*, 1992, p. 59).

CELF-Preschool composite scales were also compared to the Performance, Verbal, and Full Scale IQ scores of the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R Wechsler, 1989). Both assessments were given to a sample of 56 children (counterbalanced) and divided into two age subgroups. One group consisted of children between the ages of 3 years, 6 months and 3 years, 11 months, and the other of children between the ages of 5 years and 5 years, 11 months. Correlations between the CELF-Preschool Receptive Language scale and the WPPSI-R Verbal and Full scale scores were .67 and .70, respectively. The CELF-Preschool Expressive Language scale showed correlations between .45 with the WPPSI-R Performance Score and .65 with the WPPSI-R Verbal Score. A similar pattern was noted for the CELF-Preschool Total Language score; the strongest correlation were .72 with the WPPSI-R Verbal score and .71 with the Full Scale score (Wiig, *et al.*, 1992, pp. 59-60).

Finally, scores on the CELF-Preschool were compared to scores on the Differential Abilities Scales (DAS; Elliott, 1990; Nonverbal Cluster, Verbal Cluster, General Conceptual Cluster) in a sample of 54 children between the ages of 3 years, 6 months and 5 years, 11 months. Across every composite scale, the CELF-Preschool was most strongly correlated with the Verbal Cluster of the DAS, .63, .64, and .70 for the Receptive, Expressive and Total Language scales, respectively. Correlations between the CELF-Preschool composite scales and the DAS General Conceptual Ability Cluster scores were .62 for both the Receptive and Expressive Language scales, and .68 for the CELF-Preschool Total Language score (Wiig, *et al.*, 1992, p. 60).

Reliability/Validity Information from Other Studies

None found.

Comments

- Internal consistency varied by both age and subtest. For the most part, subtest coefficients were in the strong range, but sometimes dropped to the moderate range. However, internal reliability coefficients were weaker for the oldest children on the Basic Concepts subtest, and for children of 5 years and older on the Sentence Structure subtest. It may be that items for these subtests are not as appropriate for older children as they are for younger. Some older children may be reaching the ceiling. Internal reliabilities for composite scores were strong across age groups and scales (including the eldest group), but coefficients for the eldest groups of children were notably lower than the younger groups of children.
- Test-retest correlations for subtest and composite scores were strong for both age groups examined. The Sentence Structure subtest showed the lowest correlations across administrations (.60 - .64), but coefficients remained in the strong range. It is worth noting that the examination of test-retest reliability did not include children of 5 years and older. Given that the internal consistencies for the CELF-Preschool subtests were generally lower for older children, it would be useful to have test-retest reliability information for this age group as well.
- Interrater reliability was high for the one subtest that it was assessed for. It was reported to be the only subtest that required judgment on the part of the administrator, thus it might be *assumed* that the more objective subtests would show comparably strong interrater reliabilities. This information is not presented.
- Validity was established for the CELF-Preschool in a number of ways. Test development included multiple steps. Content validity, item discrimination, and developmental placement were considered in great detail during the course of this process and included both pilot testing and tryout testing prior to standardization. The authors engaged in extensive consultation with experts, to help assure that items were assigned to the most relevant subtests, and that the subtests contained content appropriate to the underlying construct. It is noteworthy, however, that no latent modeling was reported. Such analyses would help confirm the organization of the content into subtests and scales.
- Other analyses cited by the authors indicate that children with diagnosed language disorders are more likely to have CELF-Preschool scores at or below one or one and a half standard deviations of the mean than are children without such a diagnosis. This provides some support for the discriminant validity of the measure. However, the CELF-Preschool appears to be much more accurate at identifying true-negatives (those children

without a language disorder) than true positives (those with a language disorder). This held using both cut-off points. Thus, caution is warranted in using this measure (or at least the criterion cited in the manual) for screening purposes.

- Correlations between the subtest of the CELF-Preschool and the CELF-R were moderate to strong, supporting the concurrent validity of the measure. As would be expected, the CELF-Preschool subtests that were direct downward extensions of those found in the CELF-R were more highly correlated than subtests that slightly differed. These comparisons were only possible for children between the ages of 5 years and 6 years, 11 months, due to the ages for which the CELF-R is appropriate.
- Correlations between the CELF-Preschool and the PLS-3 were strong. The CELF-Preschool Total Language scale showed the highest overall correlations with PLS-3 scores. The correlation of the CELF-Preschool Receptive Language composite scale was higher with the comparable PLS-3 scale (Auditory Comprehension) than the PLS-3 Expressive Communication scale This was not true for the CELF-Preschool Expressive Language scale, which was equally correlated with the PLS-3 expressive and receptive language scales.
- Comparisons between the CELF-Preschool and the WPPSI-R showed moderate to strong correlations between the two measures, varying by the comparability of the scales compared. As would be expected, CELF-Preschool composite scales were more strongly related to the WPPSI-R Verbal scale than the Performance scale. The age range used in the comparisons excluded children 6 years old and older. Similarly, comparisons between the CELF- Preschool and the DAS found strong correlations between the two measures, with particularly strong correlations between the CELF-Preschool composite scales and Verbal Cluster of the DAS.

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

In a descriptive study of language ability in low-income samples in the United Kingdom, Locke, Ginsborg, and Peers (2002) assessed 223 children using the CELF-Preschool (UK edition). The children's CELF-Preschool scores were then compared to the published test norms to assess whether the scores of low-income children systematically differed. They found that children within the low-income population generally fell between moderate language delay and non-delay (one standard deviation below the mean). A majority of children in the sample fell below the expected ability for their age. It was also noted that girls scored significantly higher on the Receptive and Total Language scores than boys in this sample.

V. Adaptations of Measure

No adaptations were found other than the UK edition that was used by Locke *et al.* (2002).

Expressive One-Word Picture Vocabulary Test (EOWPVT)

I. Background Information

Author/Source

Source: Brownell, R. (2000a). *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 the author

"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, 2000a, 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, 2000a, 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 (e.g., "What are these," "what word names all of these?"). Pointing cues may be given if the child is not attending to the feature of the picture that is intended.

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, 2000a, p. 63).

Split-Half Reliability
Split-half coefficients ranged from .96 to .99, with a median of .98 (see Brownell, 2000a, 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, 2000a, 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 (Brownell, 2000a, pp. 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 between the two scorings (Brownell, 2000a, 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 (Brownell, 2000a, pp. 65-66).

Validity Information from Manual

Concurrent Validity

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

Construct Validity

A number of different findings were reported by Brownell in support of the construct validity of the EOWPVT.

- 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 (Brownell, 2000a, p. 73).
- Correlations of scores on the EOWPVT and on 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, were 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; Brownell, 2000a, p. 73).
- Corrected correlations were examined between the EOWPVT and five broad measures of language focusing on "language in connected discourse," including 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 (Brownell, 2000a, 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 found to be 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 (Brownell, 2000a, 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 (Brownell, 2000a, p. 76).
- The correlation between scores on the Expressive and Receptive One Word Vocabulary Test (uncorrected) was .75 (Brownell, 2000a, 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. (2000c). *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).

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.

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 the authors

"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; 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. 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?
For the Expressive Vocabulary subtest, the examiner presents a series of pictures, and the child is asked to give the names of pictured objects verbally. 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). There are also designated stopping points based on age. However, if the child passes all the items in the last unit intended for his or her age, additional items are administered until one is missed.

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, 1976) and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI; Wechsler, 1967; see Kaufman & Kaufman, 1983b, p. 113).

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

Predictive Validity

K-ABC Achievement Scores were highly predictive of measures of cognitive development. However, the Expressive Vocabulary subtest was not examined separately (see Kaufman & Kaufman, 1983b, p. 121).

- With 11 months between test administrations, the correlation of K-ABC Achievement Scores and PIAT (Markwardt, 1998) total scores in a sample of 29 normally developing children was .72.
- With 10 months between test administrations, the correlation between K-ABC Achievement Scores and PIAT total scores in a sample of 30 Navajo children was .82.
- With 7 months between test administrations, the correlation of K-ABC Achievement Scores with PIAT total scores in a sample of 29 "educable mentally retarded" children was .67.
- With 11 months between test administrations, the correlation of K-ABC Achievement Scores with Woodcock-Johnson (WJ-R; Woodcock & Johnson, 1989) preschool cluster scores in a sample of 31 normally developing children was .73.

- With 6 months between test administrations, the correlation of K-ABC Achievement Scores with Iowa Test of Basic Skills Vocabulary subtest scores in a sample of 18 normally developing children was .88.
- With 12 months between test administrations, the correlation of K-ABC Achievement Scores with California Achievement Test (CAT5; 1992) Total Language Battery scores in a sample of 45 normally developing children 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, 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).
- 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.

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 the author

"[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.
  - *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. 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, 2000a) 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, pp. 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, pp. 75-77).

Reliability/Validity Information from Other Studies

- Feldman *et al.* (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

Non-English Language Versions

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).

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 the authors

"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 and a series of criterion measures were examined in samples of varying ages. These correlations ranged from .62 to .91. Specifically, reported correlations between scores from criterion measures and scores on PPVT-III Form A and B scores, respectively, 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)

No studies were found examining children's PPVT-3 scores in relation to environmental variation; however, studies utilizing the PPVT-R have been published. Two such studies are McCartney (1984) and Peisner-Feinberg and Burchinal (1997).

- 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.

Preschool Language Scale – Fourth Edition (PLS-4)

I. Background Information

Author/Source

Source: Zimmerman, I. L., Steiner, V. G., & Pond, R.E. (2002). *Preschool Language Scale Fourth Edition: Examiner's Manual*. San Antonio, TX: The Psychological Corporation.

Publisher: The Psychological Corporation
19500 Bulverde Road
San Antonio, TX 78259
Phone: 1-800-872-1726
Website: www.psychcorp.com

Purpose of Measure

As described by the authors

"The Preschool Language Scale – Fourth Edition (PLS-4) is a revision of the Preschool Language Scale – Third Edition (PLS-3). The PLS-4 is an individually administered test used to identify children who have a language disorder or delay"(Zimmerman, Steiner, & Pond, 2002, p. 2). Along with providing two core language subscales (i.e., Auditory Comprehension and Expressive Communication), the PLS-4 also provides supplemental assessments: the Language Sample Checklist, the Articulation Screener, and the Caregiver Questionnaire. The set of instruments thus includes both norm and criterion referenced information about child language.

Population Measure Developed With

The authors state that one of the key reasons the PLS-3 was updated was a population shift from the standardization sample used in that version (i.e., Census data from 1980). The standardization sample of the current version (PLS-4) was selected based on demographic information obtained from the 2000 Census, and the sample was stratified by parent education, geographic region, and race.

- The sample included 1,564 children between the ages of 2 days to 6 years, 11 months.
- Age groups were created for every two months of child age (e.g., children 0 years. 0 months through 0 years, 2 months) until the age of one, at which point children were grouped in six-month spans (e.g., children 1 year, 0 months through 1 year, 5 months, 1 year, 6 months through 1 year, 11 months). There was an approximately equal distribution of boys and girls within each age group.
- The sample was comparable to 2000 Census data for regional (Northeast, North Central, South, and West) and racial distributions (white, black, Hispanic, other).
- Education levels of primary caregivers closely matched 2000 US Census education levels for parents with children between birth and 6 years, 11months of age. Ninety-four percent of the time mothers were the primary caregiver; fathers were the primary caregiver in 6 percent of the sample.
- Fifty-five and a half percent of children were listed as spending most of the day "at home with family," with the rest of the children distributed across participation in daycare,

preschool, kindergarten, 1st grade, other, or care at home with a sitter. Of the children not yet in school, 34 percent were in some form of non-familial care.

- A large majority of children in the sample spoke only English; 3.4 percent spoke both English and Chinese, and 3.4 percent spoke English and Spanish.
- Small percentages of the sample were found to have an identified condition or diagnosis, including Developmental Delay, Articulation Disorder, Prenatal Condition, and Language Disorder. Summarizing across 10 diagnoses/conditions and an "other" category, 13.2 percent of the children in the sample were found to have one or more of the diagnoses/conditions.

Age Range Intended For

Birth through 6 years, 11 months.

Key Constructs of Measure

The PLS-4 has two core subscales, Auditory Comprehension and Expressive Communication, a composite Total Language scale, and three supplemental assessments, the Language Sample Checklist, the Articulation Screener, and the Caregiver Questionnaire.

- *Auditory Comprehension.* This subscale measures how much language the child understands. For infants and toddlers this subscale involves precursors of language (e.g., attention to speakers, responding to basic requests like "no-no"). Preschool tasks address concepts such as comprehension of vocabulary and grammar rules. Tasks for older children involve comprehension of complex sentences and ability to make inferential decisions.
- *Expressive Communication.* This subscale measures how well the child communicates with others. Like the Auditory Comprehension subscale, Expressive Communication tasks vary by child age. Infant and toddler tasks initially assess rudimentary aspects of expressive language, such as the ability to make sounds of pleasure, and later involve tasks that require the child to demonstrate verbally an understanding of language concepts such as plural tense. Preschool-age tasks focus on more advanced aspects of expressive language, such as naming items and actions and using proper sentence structures. Tasks for the oldest children for whom the PLS-4 is designed involve still more complex aspects of expressive communication, such as storytelling.
- *Total Language.* The Total Language score is the sum of the Expressive Communication and Auditory Comprehension standard scores.
- *Supplemental Assessments*
    - *Language Sample Checklist.* This can be used with any child who speaks in connected utterances. It is designed to provide an overview of utterance content and structure.
    - *Articulation Screener.* This assessment can be used with children from 2 years, 6 months through 6 years, 11 months of age. It is used as a screener for possible language deficits to determine whether the child should be referred for further in-depth assessment. The screener is criterion scored.
    - *Caregiver Questionnaire.* This part of the PLS-4 is intended to elicit information from the child's caregivers (i.e., mother, father, other) about the child's behavior at home. It may be used when assessing children under 3 years of age. Information obtained from this questionnaire may provide information to

facilitate interpretation of Expressive Communication and Auditory Comprehension subscales, but does not replace them.

Norming of Measure (Criterion or Norm Referenced)
Norm referenced, with the exception of the criterion referenced Articulation Screener.

Comments

- The use of the identified primary caregiver's level of education, (rather than maternal education in all cases) as a stratification variable in the standardization sample is noteworthy. It is not clear if the proportion of children with father as the primary caregiver is representative of the U.S. population, or whether having the father as the primary caregiver affects language development.
- It should be noted that reporting the main care environment for children in the standardization sample is rarely done in reporting on measures development. Child care participation was not used as a stratification variable, but given findings linking child care quality to language development (e.g., NICHD Early Childcare Research Network, 2000b; Burchinal, Roberts, Riggins, Zeisel, Neebe, & Bryant, 2000; McCartney, 1984; Vernon-Feagans, Emanuel, & Blood ,1997), its inclusion in sample description for this measure provides unique and useful information.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
Child tasks vary depending on the construct being assessed and the difficulty level of the task.

- For Auditory Comprehension, early tasks include whether the child glances at the person speaking to him or her; tasks for older children require more involved responses such as appropriate play with an object (e.g., bouncing a ball, stacking blocks, making a car "go"). Other behaviors include pointing responses to verbal cues and more complicated actions such as arranging pictures according to the way the pictured subjects rhyme.
- Expressive Communication tasks include assessing young infants' ability to suck/swallow (e.g., baby can suck pacifier) or to vary the pitch, timbre, or length of a cry. Later behaviors include the child's ability to imitate words articulated by the assessor, the ability to name toys, and the degree to which a child can provide verbal answers to questions about hypothetical situations). More complicated tasks require the child to understand and correct grammatical errors.
- Both the Auditory Comprehension and Expressive Communication subscales have several different ways of scoring a correct response, depending upon child age and the type of question. Especially when assessing children in the younger age ranges, the assessor may score an item as correct if the target behavior is elicited from the child in response to a specific stimulus (scored E); whether the child exhibits the target behavior at any time during the testing period, irrespective of response to a specific stimulus (scored S); or if, without elicited response, the caregiver can provide detailed examples of

the child doing such behaviors at home (scored C). This three-way scoring option varies by question, thus correct responses can sometimes include all three, E and S, or just E.

- Both subscales use basals and ceilings to establish where the an assessment should begin and end. Initial starting points for each subscale are based on child age, and a basal is established when the child passes three consecutive tasks. The ceiling has been reached when the child receives a score of zero on seven consecutive tasks.

Who Administers Measure/Training Required?

Test Administration

Speech-language pathologists, early childhood specialists, psychologists, educational diagnosticians, properly trained paraprofessionals, and others who have experience and training in assessment can administer the assessment.

Data Interpretation

The test should be interpreted by speech-language pathologists, early childhood specialists, psychologists, educational diagnosticians, and clinicians who have experience and training in diagnostic assessment.

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

This assessment is administered one-on-one or with family members present. The PLS-4 materials provide information and suggestions for giving the assessment in different locations (e.g., clinic or home), as well as ways to use the presence of others during the assessment (e.g.,, parent, siblings), while still providing the assessment in the required manner.

Time Needed and Cost

Time

Administration time for the PLS-4 varies by the age of the child being assessed. Times for children from birth to 11 months average between 20 and 40 minutes, times for children from 12 months to 3 years, 11 months average between 30 and 40 minutes, and times for children between the ages of 4 years and 6 years, 11 months, average between 25 and 45 minutes.

Cost

- PLS-4 English Value Pack with Manipulatives: $235.00
- PLS-4 English Basic Kit: $185.00
- Examiner's Manual, English: $60.00

Comments

- The Examiner's Manual provides numerous suggestions for administering the PLS-4 to children with special needs (e.g., children with developmental delays, children with sensory impairments, autistic children, children who use sign language), including what to expect, what modifications can be made to the measure, ways to set varying basals, and alternative ways of scoring. However, the validity and reliability of the PLS-4 for children with special needs is unclear.
- The flexibility to include parents and siblings, and to conduct assessments in differing environments, may be both a strength and a weakness of the PLS-4. On the positive side, it may allow more accurate readings for temperamentally introverted children and reduce

the minimum age for assessment. The authors argue that caregiver participation is necessary for proper administration with very young children. One could also argue, however, that nonstandard administration may increase the possibility of assessment errors.

- It is unclear how the various ways of scoring (i.e., E, S, and C) might affect the reliability and validity of the measure. The determination of which tasks included alternate scoring options (as opposed to scoring elicited behavior only) was based on a task by task analysis of norming sample data that assessed the degree to which caregivers provided examples of target behavior, the degree to which examiners consistently scored the caregiver's information correctly, and the degree to which spontaneous behavior was consistently noted for a task.

III. Functioning of Measure

Reliability Information from the Manual[10]

Test-retest reliability

A randomly selected subsample of 218 children (117 females) drawn from the standardization sample was used for an analysis of test-retest reliability. The sample consisted of children between the ages of 2 years and 5 years, 11 months, with a mean age of 3 years, 5 months. The time between test and retest ranged from two to fourteen days, with a mean of 5.9 days; the same examiner assessed each child on both occasions. The authors report test-retest correlations within 6-month age bands.

- For Auditory Comprehension, correlations ranged from .83 for the 5 years to 5 years, 5 months age group, to .95 for the 2 years, 6 months through 2 years, 11 months age group.
- Test-retest correlations for Expressive Communication ranged from .82 for the 2 years through 2 years, 5 months age group, to .95 for the 4 years, 6 months through 4 years, 11 months age group.
- Total Language test-retest correlations ranged from .90 for the 2 years through 2 years, 5 months age group to .97 for the 2 years, 6 months through 2 years, 11 months age group (Zimmerman *et al*., 2002, p.188).

Internal consistency

Data from the standardization sample were used to assess the internal consistency of the PLS-4. Coefficient alphas were presented for 3-month age bands from 0 to 11 months, and for six month age bands from age 1 year through 6 years, 11 months (see Zimmerman *et al.*, 2002, p. 189).

- Alphas for the Auditory Comprehension subscale ranged from .66 for the oldest age group (6 years, 6 months through 6 years, 11 months) to .94 for children ages 3 years through 3 years, 5 months. The alpha for the full sample was .86.
- Expressive Communication subscale alphas ranged from .73 for the 9 months through 11 months age group to .95 for the 3 years, 6 months through 3 years, 11 months age group. The alpha for the full sample was .91.

[10] Section III is reported based on the Auditory Comprehension and Expressive Communication subscales and the Total Language scale, as psychometric information for the supplemental scales is not readily available.

- Total Language composite scores ranged from .81 for the 9 months to 11 months age group to .97 for both the 3 years through 3 years, 5 months and the 3 years, 6 months through 3 years, 11 months age groups. The alpha for the full sample was .93.

Interrater reliability
The authors indicated that while most PLS-4 tasks are objectively scores, some Expressive Communication tasks were open to scorer interpretation. Interrater reliability was assessed for these tasks. Fifteen examiners, most of whom were elementary education teachers, were trained in PLS-4 scoring rules and were allowed 3 weeks of practice with test protocols. Following this training period, a random sample of 100 test protocols from the standardization sample (an average of 6 tasks per 3- or 6-month age group) were each scored by two of these scorers, working independently. Agreement for these tasks was 99 percent, and the correlation between Expressive Communication scores was .99 (see Zimmerman *et al.*, 2002, p. 190).

Validity Information from the Manual
Zimmerman *et al.* (2002) present validity information in a different way than in most other manuals, following recommendations found in the *Standards for Educational and Psychological Testing* developed by a joint committee of the American Educational Research Association, the American Psychological Association, the National Council on Measurement in Education, and the American Research Association (1999). Rather than focusing on different forms of validity (e.g., construct, discriminant, predictive), discussion of PLS-4 validity is organized in terms of different "…types of evidence…[that] support the test's interpretations and uses" (p. 190).

Evidence Based on Test Content
Zimmerman *et al.* (2002) state that the "PLS-4 offers a thorough and balanced sample of language behaviors" (p. 190). To ensure that this was the case, several procedures were followed.

- A comprehensive literature review was carried out to assure that important language development milestones were covered by the PLS-4. The review noted how assessment procedures had been developed in previous work to elicit key language behaviors. In addition, content covered in Language Arts curricula was reviewed to identify aspects of language required of 5- and 6-year old children in school.
- Clinicians who used the PLS-3 (the previous version of the PLS) were asked for feedback on possible improvements to the measure, and many of these revisions were included in the PLS-4.
- Speech-language pathologists working in a range of clinical settings developed new tasks for the measure.
- A panel of experts in assessment, cultural/linguistic diversity, and/or regional language issues was consulted to assess possible biases in the PLS-4 tasks, stimuli, and items.
- Based on this work, a version of the PLS-4 was then developed and piloted with a group of children similar to those in the standardization sample. The pilot data were then reviewed statistically, and experts completed qualitative reviews of the tasks and items. Tasks or items found to be biased were modified or dropped

The PLS-4 Examiner's Manual offers a great deal of further detail about the scope of developmental language indicators the measure covers, including precursors, semantics,

structure, morphology, integrating skills and phonological awareness (Zimmerman *et al.*, 2002, pp.190- 207).

Evidence Based on Response Processes

The authors provided evidence that the PLS-4 measures the language skills that it is designed to measure, rather than other skills or abilities not directly related to language. During the PLS-4 development phase, task selection involved several considerations related to response processes. Tasks were reviewed to ensure that 1) they focused on intended skills, 2) they did not require response skills that were outside of the range of abilities of children at a given age, 3) they did not require additional abilities that could confound results (i.e., memory abilities not directly related to language skills), and 4) task content was interesting to children. To assess whether children's responses on the tasks were reflecting the intended language processes, a small group of 4- to 6-year old children was asked to explain why they gave particular answers during testing. Their responses to this question were evaluated and in some cases items were dropped or modified. In addition, examiners who participated in a Tryout phase of PLS-4 development were given a questionnaire designed to elicit information regarding items that were difficult to administer or to score. Problematic items were deleted from the final version of the PLS-4. The authors assert "the evidence supports the hypothesis that the desired response processes are being delivered" (Zimmerman *et al.*, 2002, p. 208). However detailed information regarding the justification for changing or deleting tasks is not provided.

Evidence Based on Internal Structure

The authors summarized data on the degree to which the tasks, individual items and subscales were related to each other, and the degree to which these relations followed expected patterns for language abilities. The authors pointed to the high internal consistency of the subscales (see reliability section) as evidence of the internal structure validity of the test. In addition, they point to the .74 correlation between the Auditory Comprehension and Expressive Communication subscales as evidence supporting the validity of the PLS-4 because "…these two subscales should have a fairly strong correlation because they both purport to measure language, but not too strong because they measure different aspects of language" (Zimmerman *et al.*, 2002, p.209).

Evidence Based on Relations to Other Variables

PLS-4 scores were compared to performance on the Denver II (Frankenburg & Bresnick, 1998) screener in a sample of 37 children (19 male, 18 female). The majority of the children were white (73 percent), 22 percent were black and 5 percent were considered "other, " and the children in this sample had parents with generally higher levels of education: 38 percent had completed some college, and 30 percent had at least a four-year degree. Each child was assessed with each measure (with the order of assessment counterbalanced) on the same day, by the same examiner. The Denver II does not give standard scores. Rather, the overall rating of children as "normal," "suspect," or "untestable" was compared to how the children scored on the PLS-4. All 37 children scored within a standard deviation of the mean on the PLS-4 and were classified as "normal" on the Denver II. In other words, all children in this sample were identified as being within normal ranges of development on both measures, which the authors took to indicate "…a high level of agreement, as hypothesized" (Zimmerman *et al.*, 2002, p. 210).

In another sample, PLS-4 scores were correlated with scores on the earlier version of the measure, the PLS-3. The sample for this analysis consisted of 104 children (57 female, 47 male), ranging in age from 2 months to 6 years, 11 months. The majority of children in the sample were white (73 percent). Parents of the children in this sample had relatively low levels of education, with 39 percent having less than a high school diploma, 37 percent a high school diploma, and 24 percent having completed 1 to 3 years of college. Each child was administered both assessments by the same examiner, between 2 days and 2 weeks apart, with order counterbalanced. The Auditory Comprehension subscale of the PLS-4 correlated .65 with the Auditory Comprehension subscale of the PLS-3. Similarly, the PLS-4 Expressive Communication subscale correlated .79 with the PLS-3 Expressive Communication subscale (Zimmerman *et al.*, 2002, p. 210).

Further studies were carried out to assess the clinical validity of the PLS-4. PLS-4 scores for typically developing children were compared with scores for children with identified impairments likely to affect language functioning. The four clinical groups included in these studies were 1) children identified as having a language disorder, 2) children having a developmental delay, 3) children diagnosed with autism, and 4) children with hearing impairment.

- *Children with a language disorder.* PLS-4 scores for 75 3-, 4- and 5- year-old children who had been diagnosed with language disorder were compared with scores for 75 children without identified language disorders drawn from the standardization sample. The two groups were matched based on age, gender, ethnicity and parent education. The children with language disorders were currently in a remediation program, had both receptive and expressive language problems, and were only included if they could take the test in the usual fashion (that is, they had no gross or fine motor problems). Sensitivity and specificity analyses were carried out to assess how well the PLS-4 scales differentiated typically developing children from those with language disorders. Sensitivity was defined as the probability that a child with a disorder would be correctly identified as having a disorder (i.e., a true positive). Specificity was defined as the probability that a child without a language order would be classified as not having a problem (i.e. a true negative). Analyses were carried out separately for the age groups, and did not vary greatly by scale or child age.
    - For children between 3 years and 3 years, 11 months, the PLS-4 Auditory Comprehension subscale showed a sensitivity of .79 (i.e., 79 percent of children with an identified language disorder were identified as having a language disorder on the basis of their PLS-4 scores) and specificity of .92 (i.e., 92 percent of children without a diagnosed language disorder were similarly identified as not having a language disorder on the basis of their PLS-4 scores). The Expressive Communication subscale had a sensitivity of .79, and a specificity of .88. The Total Language score had a sensitivity of .83 and a specificity of .88 (see Zimmerman *et al.*, 2002, p. 213).
    - For children between the ages of 4 years and 4 years, 11 months the PLS-4 Auditory Comprehension subscale showed a sensitivity of .78 and a specificity of .96. The Expressive Communication subscale had a sensitivity of .74 and a specificity of .91. The Total Language score had a sensitivity of .78 and a specificity of .96 (see Zimmerman *et al.*, 2002, p. 213).

  - For children between 5 years and 5 years and 11 months, the PLS-4 Auditory Comprehension subscale showed a sensitivity of .82 and specificity of .89. The Expressive Communication subscale had a sensitivity of .79, and a specificity of .75. The Total Language score had a sensitivity of .79 and a specificity of .82 (see Zimmerman *et al.*, 2002, p. 214).
  - Across all ages, the PLS-4 Auditory Comprehension subscale demonstrated a sensitivity of .80 and specificity of .92. The Expressive Communication subscale had a sensitivity of .77 and specificity of .84. The Total Language score demonstrated values of .80 for both sensitivity and specificity (see Zimmerman *et al.*, 2002, p. 214).
- *Children with Developmental Language Delays.* A total of 116 children (58 with language delays and a matched group of 58 typically developing children) ranging from 12 months through 2 years, 11 months of age (with a mean age of 2 years, 2 months) were assessed with the PLS-4. The language delayed children in this study had been diagnosed as demonstrating a language delay by an agency where they were receiving services; these children were enrolled in remediation programs at the time that they were assessed with the PLS-4. Based on earlier testing, these children were identified as demonstrating delays in language skills of six to 10 months. Children identified as having language delays had mean standard scores on the PLS-4 Auditory Comprehension, Expressive Communication, and Total Language of 78.7, 75.8, and 75.0, respectively. In contrast, mean standard scores for the typically developing children in this study were 101.8, 100.7, and 101.5 for Auditory Comprehension, Expressive Communication, and Total Language, respectively). No significance tests were reported for group comparisons (see Zimmerman *et al.*, 2002, p. 215).
- *Children with Autism*. Forty-four children (32 male, 12 female) diagnosed with autism were matched with 44 typically developing children from the norming sample on the basis of age, gender, ethnicity, and parent education. The children were predominantly white and had relatively well educated caregivers (39 percent with some college, 22 percent with four years of college or more). The children ranged in age from 3 years, 6 months to 6 years, 7 months. All of the children with autism were enrolled in programs for children with autism at the time the PLS-4 was administered. Mean scores for the Auditory Comprehension, Expressive Communication and Total Language scales were compared. Across scales, typically developing children did better than children with autism. Standard score means for the autistic group were 66.8, 65.5, and 64.4 for Auditory Comprehension, Expressive Communication, and Total Language scores, respectively. In contrast, standard score means for the matched control group of typically developing children were 102.8, 102.3, and 103.0 for Auditory Comprehension, Expressive Communication, and Total Language, respectively. No significance tests were reported for these group comparisons (see Zimmerman *et al.*, 2002, p. 216).
- *Children with Hearing Impairments*. Thirty-two children (36 male, 28 female) diagnosed with moderate to severe hearing impairments were matched with 31 typically developing children from the norming sample on the basis of age, gender, race, and parent education. Sixty-nine percent of the children were white, 13 percent were Hispanic, and 9 percent were black. Parent education was relatively high for this sample; 37 percent of parents had some college, and 34 percent had a 4-year college degree or more. Children ranged in age between 3 years, 6 months and 6 years, 11 months, with a mean age of 5 years, 5

months. Sixty percent of the children with hearing impairments used sign language exclusively, and they had all been diagnosed as having expressive/receptive language disorders and were enrolled in remediation programs. Examiners presented the stimuli in sign language or a combination of sign language and spoken words, depending upon the child's needs. The children with hearing impairments demonstrated lower standard scores on the PLS-4 than did the matched group of typically developing children. Standard score means for hearing impaired children were 70.7, 66.7, and 66.5 for Auditory Comprehension, Expressive Communication and Total Language scores, respectively. In contrast, standard score means for Auditory Comprehension, Expressive Communication, and Total Language for the matched group of typically developing children were 109.3, 109.0, and 110.3, respectively. The authors further indicate that children with severe impairments had difficulty across domains of language development with the exception of social communication, while those with moderate hearing impairments had more specific difficulties. Generally, the children with hearing impairments had a wide range of skills involving semantic relationships and syntactic constructions, but had difficulty communicating complex, sequential constructions using proper grammar. No statistical tests of differences between hearing impaired and typically developing children were reported (Zimmerman *et al.*, 2002, p. 217).

Reliability/Validity Information from Other Studies

Because the PLS-4 is a relatively new version of the PLS, little independent psychometric study has been carried out on it. In a study with an earlier version of the PLS, Mott (1987) found that Total Language Scores on the PLS-R were strongly correlated with the Battelle Developmental Inventory (BDI; Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1984) language scales. In that study, correlations between BDI Receptive and Expressive Developmental Quotients and PLS-R Total Language Scores were .54 (n.s.) and .75, respectively. The strongest association reported by Mott was .81, between BDI Total Communication scale scores and PLS-R Total Language Scores. This illustrates some concurrent validity between the older version of the measure and another measure of language development. It should be noted, however, that not only has the PLS been revised since this study, but the sample was small (20), all white, and consisted of children with language impairments.

Comments

- Test-retest reliability for the PLS-4 appears to be quite strong.
- Internal consistency ranged from moderate to strong for the Auditory Comprehension subscale and was consistently strong for the Expressive Communication subscale. There was variation across child age subgroups in the internal consistency coefficients, but these did not fall into a consistent pattern (either increasing or decreasing with child age).
- Interrater reliability was very strong, but analyses were only presented for the subsection of the Expressive Communication subscale that necessitated open-ended interpretation of child behavior. It is worth noting that interrater reliability was assessed in a sample of teachers who were trained in PLS-4 administration, but who had no previous experience or training in psychological or cognitive assessment. used reliably by
- It does not appear that any analysis was done examining the reliability of the different means of recording "correct" responses (i.e., elicited directly, happened anytime during testing, caregiver report). The possibility exists that different assessors differ in the extent

to which they notice child behavior outside the elicited responses, or evoke responses from the caregiver. In general, none of the reliability analyses provided in the manual can address the extent to which children would receive the same scores from two different evaluators conducting independent assessments.

- Some of the strongest evidence supporting the validity of the PLS-4 comes from the authors' extensive literature review, expert consultation, and piloting. An iterative process of working with experts in speech pathology and special populations while piloting tasks to assess their utility was used to establish construct validity. Although quantitative methods such as factor analysis or IRT were not reported, it appears that the process followed in constructing the measure resulted in a high level of face validity.
- The authors did not assess concurrent validity with other language assessments with overlapping constructs (e.g., PPVT, TELD, EWOPVT); rather, they did compared PLS-4 results with results from a screening test, the Denver II. The utility of this comparison is limited by the sensitivity and specificity of the Denver II. It is further limited by the fact that all children were within normal ranges on both tests.
- The analyses involving comparisons of typically developing and matched exceptional children are noteworthy. The PLS-4 differentiated those with language and developmental disorders from those whose language was developing typically. Specificity and sensitivity were high across age groups.

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

The PLS-4 is a relatively new version of the Preschool Language Scales and no studies were found using it as a child outcome measure in relation to environmental variation. The previous version of the measure, the PLS-3 (Zimmerman, Steiner, & Pond, 1992), was used by the NICHD Early Child Care Research Network (e.g., 2002a). This longitudinal study used 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, 15, 24, 36, and 54 months of age. The PLS-3 was administered at 54 months. Higher quality care, quality care that improved over time, and more time in center care (as opposed to other non-maternal care) were predictive of higher PLS-3 scores even after controlling for a range of family, parenting, and economic factors. In another analysis of the same sample, both PLS-3 subscales were used along with other measures as indicators of a latent construct, cognitive competence (NICHD Early Childcare Research Network, 2002b). Variation in structural features of child care such as ratio and caregiver training predicted child cognitive competence. Further, this relationship was mediated by variations in children's immediate experiences in child care, including quality of caregiver-child interaction.

V. Adaptations of Measure

Spanish Version of PPVT-III

A Spanish version of the PLS-4 is available.

Reynell Developmental Language Scales: U.S. Edition (RDLS)

I. Background Information

Author/Source

Source: Reynell, J. & Gruber, C. P. (1990). *Reynell Developmental Language Scales – U.S. Edition.* Los Angeles, CA: Western Psychological Services.

Publisher: Western Psychological Services
12031 Wilshire Blvd.
Los Angeles, CA 90025-1251
Phone: 1- 800-648-8857
Website: http://www.wpspublish.com

Purpose of Measure

As described by the instrument publisher

"The Reynell Developmental Language Scales simplify what is often a difficult task—measuring language skills in young or developmentally delayed children. Widely known for clinical usefulness, these scales assess two processes essential to language development: verbal comprehension and expressive language" (Western Psychological Services, 2002, https://www-secure.earthlink.net/www.wpspublish.com/Inetpub4/catalog/W-275.htm).

Population Measure Developed With

The American standardization sample included 619 children assessed at nine sites in eight states. Children ranged in age from 1 year to 6 years, 11 months. Child gender was distributed approximately evenly, with 49 percent males and 51 percent females. The sample was stratified to be nationally-representative, based on 1987 U.S. Census data for adults ages 25 to 44, for geographic region (east, north-central, south, and west), ethnicity (white, black, Hispanic and Asian), and parent education (less than high school, high school graduation, some college, and college graduation or higher). Approximately 84 percent of the sample was white, 10 percent was black, 5 percent was Hispanic, and 2 percent was Asian. Approximately 34 percent of parents had a four-year college degree or higher education, 18 percent had some college education, 32 percent had graduated from high school, and 16 percent had less than high school graduation.

Age Range Intended For

Ages 1 year to 6 years, 11 months.

Key Constructs of Measure

There are two primary scales derived from the Reynell Developmental Language Scales (RDLS): Verbal Comprehension and Expressive Language.

- *Verbal Comprehension*. This scale taps receptive language abilities. There are 67 items organized into 10 sections representing successively more developmentally advanced receptive language abilities. The earliest abilities involve verbal preconcepts, indicated by infants' differential responsiveness to familiar words (e.g. "Daddy") or phrases

("Daddy's coming") spoken by the mother. The most advances abilities assessed by the RDLS involve verbal reasoning and drawing inferences.

- There are two versions of the Verbal Comprehension scale—VCA and VCB. VCA should be used for most children; VCB was developed to allow testing of children with severe speech and motor impairments for whom eye-pointing may be the only method of response. The authors further suggest that VCB may be used with very shy or withdrawn children, because "The Stimulus Materials are of such intrinsic interest to children that few can resist at least looking at the objects in response to an item, and that in itself usually constitutes a selective response" (Reynell & Gruber, 1990, p. 38). The first halves of the two versions are identical; the primary difference is that the number of objects per test item is reduced for items on the second half of the test, in order to make it possible to code eye-pointing accurately.
- *Expressive Language.* The Expressive Language scale includes 67 items. There are three subscales tapping different aspects of early language development. The three subscales, Structure, Vocabulary, and Content, are ordered developmentally with respect to the emergence of the language abilities they tap; however there is substantial overlap in the continuing development of each aspect of expressive language.
    - *Structure.* This subscale is designed to examine children's spontaneous vocalizations, ranging from early pre-symbolic vocalizations (e.g., single-syllable sounds, babbling, intonation patterns) to complex sentences with subordinate clauses. If there is not sufficient spontaneous speech to code all items on this subscale, simple scenarios may be used to elicit a larger language sample.
    - *Vocabulary.* This subscale is broken into three sections that explore different types of vocabulary and stimuli, including 1) naming objects, 2) identifying objects and events from pictures, and 3) defining words, ranging from object words to concepts, without any visual representation.
    - *Content.* This subscale is used to measure creative uses of language involving the ability to verbalize connected ideas.

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

Comments
The RDLS was originally developed to meet the needs of clinicians working in a center for handicapped children in Great Britain. The measure was subsequently revised and standardized in Great Britain. The United States version includes some minor changes in language due to differences in language usage conventions, as well as some changes in test administration procedures to adjust for differing testing conventions in the two countries.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

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

What the child is asked to do varies by construct and the subtest being administered.

- One Expressive Language subscale, Structure, primarily involves the examiner observing and coding spontaneous vocalizations and speech, although additional speech samples may be elicited from older children through the use of simple scenarios that the child is asked to talk about. The Vocabulary and Content subscales require the child to provide vocal responses, including identifying stimuli that are physically present or that are pictured on cards. Items become increasingly difficult and require more developmentally advanced language and cognition skills as testing continues.
- The Verbal Comprehension scale measures receptive language; the child is not required to provide verbal responses but rather is required to indicate through physical actions that he or she has some level of understanding of words and phrases spoken by the parent (for very young children) or by the examiner. For the youngest children, codable responses include physical evidence that words like "Daddy" are differentiated from less affectively-laden words (i.e., that infants respond differentially to words that have important meaning in their lives). Beyond this level, children are asked by the experimenter to perform actions including 1) pointing to objects, pictures of objects, or figures (animals and people), 2) associating two objects together as requested by the examiner, 3) pointing to objects that are identified by the activities that can be performed with them, 4) pointing to figures that are identified by the activities they perform, 5) pointing to objects or figures, or physically moving them as directed by the examiner, based on color, size, or location, including objects identified negatively (i.e., Which one is *not*…), and 6) pointing to dolls to answer questions that require inferences.
- As indicated above, the standard version of the RDLS Verbal Comprehension scale (VCA) requires children to point to objects or pictures of objects, or to perform simple actions as requested by the examiner. The alternative version (VCB) was developed to test the receptive language skills of children who cannot or will not respond motorically. This version replaces physical pointing or manipulation of objects with eye pointing.

Who Administers Measure/Training Required?

Test Administration

According to the authors, "Users of this test should be trained in the application of individually administered clinical instruments to young or developmentally delayed children" (Reynell & Gruber, 1990, p.1).

Data Interpretation

No differences in qualifications for administration and interpretation are discussed.

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

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

Time Needed and Cost

Time

Less than 30 minutes (total for both subscales).

Cost

- Full kit $499.00, includes complete stimulus set, manual, scoring sheets and carrying case.
- Manual $60.00.

III. Functioning of Measure

Reliability Information from the Manual

Internal Consistency

Spearman-Brown corrected split-half reliability coefficients were calculated for 6-month age groups between 1 year and 6 years, 11 months (Reynell & Gruber, 1990, p.49).

- Reliabilities for Verbal Comprehension (VCA) differed somewhat by age. Reliabilities were above .90 (ranging from .90 to .93) for children between the ages of 1 year, 6 month and 3 years, 5 months of age (4 age groups). Reliabilities were between .80 and .87 for the youngest age group (1 year through 1 year, 5 months) and for children ages 3 years, 6 months through 5 years, 5 months (4 age groups). For the three age groups including children ages 5 years, 6 months through 6 years, 11 months, reliabilities ranged from .72 to .78.
- The same pattern of age differences in reliabilities was evident for VCB, the alternate Verbal Comprehension scale. For this scale, reliabilities for children in the four age groups spanning 5 years through 6 years, 11 months ranged from .70 to .75; reliabilities for children ages 1 year through 1 year, 5 months and 3 years, 6 months through 4 years, 11 months (3 age groups) ranged from .80 to .83; and reliabilities for children ages 1 year, 6 months through 3 years, 5 months (4 age groups) ranged from .90 to .92.
- Expressive Language split-half reliabilities were calculated for children ages 1 year, 6 months and older. These reliability coefficients followed the same general age pattern as for Verbal Comprehension, but to a lesser extent. Reliabilities for children ages 1 year, 6 months through 1 year, 11 months, and 2 years through 2 years, 5 months were .90 and .93, respectively. Coefficients for children ages 2 years, 6 months through 4 years, 11 months (5 age groups) ranged from .85 through .88. Coefficients for children ages 5 years through 6 years, 5 months (3 age groups) ranged from .74 to .76. Finally, the coefficient for children ages 6 years, 6 months through 6 years, 11 months was .80.

Validity Information from the Manual

Content Validity

The authors state that the content validity of the RDLS is related to "…whether test questions focus on language skills and whether coverage within that domain is sufficiently broad to warrant interpretation based on RDLS results" (Reynell & Gruber, 1990, p. 51). The RDLS was originally developed in a clinic setting; items were initially selected from existing intelligence and developmental assessments and additional items were developed that were thought to represent different levels in language development. This experimental version was subsequently revised and standardized on a British sample; further revisions were conducted and the test was again standardized on an American sample for use in the U. S. Chapter 5 of the manual (pp. 36-39) discusses the types of items found within each section, providing theoretical rationales for each section based on developmental levels in language functioning thought to be necessary for

successful completion of items within each section. The authors state that test sections are "…arranged in the developmental sequence suggested by clinical trials and standardization, but there is no rigid developmental sequence for a given child, and the processes inevitably involve some degree of overlapping interrelatedness" (Reynell & Gruber, 1990, p.37).

Construct Validity

Reynell and Gruber (1990, p. 51) describe the central issue relevant to the construct validity of the RDLS as "…whether language development represents a domain that is sufficiently clear and focused to permit assessment and whether the results of such an assessment have the characteristics expected of language development." The authors cite the results from the internal reliability section as evidence of construct validity, noting that large split-half reliability coefficients are possible "…only when individual test items tend to be highly related to a single underlying dimension" (p. 51). As evidence that the RDLS assess language development, the authors also point to the fact that, within the standardization sample, the mean number of items responded to correctly increases steadily across 3-month age bands for all scales (see p. 47).

Concurrent Validity

Concurrent validity was examined in a New Zealand sample of 225 children who were seen in a clinic setting. These children were assessed with the British version of the RDLS within a month of their fourth birthday, and again a year later. Approximately half of the sample (118) had experienced some type of neonatal problem. In addition to being tested with the RDLS, the 4-year-old assessment, which was conducted over several days, included the Pictorial Test of Intelligence (PTI; French, 1964) and the Stanford-Binet Intelligence Scale (Terman & Merrill, 1960). At the age 5 assessment, all children were assessed with a battery of tests that included the RDLS, the Illinois Test of Psycholinguistic Ability (ITPA; Kirk, McCarthy, & Kirk, 1968), the Wechsler Preschool and Primary Scale of Intelligence (WPPSI; Wechsler, 1967), the Predictive Test of Articulation (Van Riper & Erickson, 1968), selected subtests from the Frostig Developmental Test of Visual Perception (Frostig, 1966), and the Goodenough-Harris Draw-a-Man Test (Goodenough & Harris, 1963). The authors predicted that correlations between RDLS scales and verbal or language scales would be higher than correlations with measures of skills within other domains. Concurrent correlations between RDLS scales and scales of the other measures are as follows (Reynell & Gruber, 1990, pp. 54-55).

- *RDLS Verbal Comprehension*
    - At age 4, correlations with PTI scales were .63, .72, and .54 with PTI Vocabulary, Information and Comprehension, and Number scales, respectively. Correlations with PTI Visual Discrimination, Visual Similarities and Visual Recall scales were .59, .55, and .57, respectively.
    - At age 5, correlations with ITPA auditory and verbal scales ranged from .53 with Verbal Expression to .74 with Auditory Association. Correlations with ITPA scales involving visual and manual abilities ranged from .43 with Manual Expression to .57 with Visual Association.
    - At age 5, correlations with WPPSI Verbal IQ subscales ranged from .59 with Sentences to .76 with Information. For WPPSI Performance IQ subscales, correlations ranged from .41 for Block Design to .60 for Picture Completion.
    - At age 5, RDLS Verbal Comprehension correlated .43 with scores on the Predictive Screen Test of Articulation.

    - At age 5, correlations with Frostig Developmental Test of Visual Perception scales Position in Space, Eye-Hand Coordination, and Figure Ground Perception were .42, .53, and .48, respectively.
    - At age 5, RDLS Verbal Comprehension scores correlated .51 with scores on the Goodenough-Harris Draw-a-Man Test.
- *RDLS Expressive Language*
    - At age 4, correlations with PTI Vocabulary, Information and Comprehension, and Number scales were .64, .62, and .48, respectively. Correlations with PTI Visual Discrimination, Visual Similarities and Visual Recall scales were .56, .52, and .52, respectively.
    - At age 5, correlations with ITPA auditory and verbal scales ranged from .53 with both Auditory Closure and Auditory Reception, to .74 with Auditory Association. Correlations with ITPA manual and visual scales ranged from .44 with Visual Perception to .54 with Visual Association.
    - At age 5, correlations with WPPSI Verbal IQ subscales ranged from .58 with Similarities to .72 with Information. For WPPSI Performance IQ subtests, correlations ranged from .40 with Block Design to .54 with Picture Completion.
    - At age 5, RDLS Expressive Language correlated .56 with scores on the Predictive Screen Test of Articulation.
    - At age 5, correlations with Frostig Developmental Test of Visual Perception scales Position in Space, Eye-Hand Coordination, and Figure Ground Perception were .36, .47, and .44, respectively.
    - At age 5, RDLS Expressive Communication scores correlated .46 with scores on the Goodenough-Harris Draw-a-Man Test.

Predictive Validity

In the New Zealand study, 185 of the 225 children seen at age 4 were reassessed three years later, at age 7, with the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949). Four–year old Stanford-Binet intelligence scores were also available and used to compare how well the RDLS and the Stanford-Binet each predicted subsequent WISC scores.

- RDLS Verbal Comprehension scale scores at age 4 correlated 71, .69, and .61, with age 7 WISC Full Scale, Verbal, and Performance IQ scores, respectively.
- The RDLS Expressive Language scale correlated .63, .67, and .49 with age 7 WISC Full Scale, Verbal, and Performance IQ scores, respectively.
- The Stanford-Binet was comparably correlated with WISC scales, with correlations of .70, .72, and .57 with WISC Full Scale, Verbal, and Performance IQ scores, respectively (Reynell & Gruber, 1990, p. 56). The authors concluded that "…the RDLS Verbal Comprehension scale and the Stanford-Binet have virtually the same very high levels of predictive validity…" (p. 53).

In another study conducted in Great Britain of 5-year-old children being treated for language delays (sample size not provided), children were assessed with the RDLS. Five years later, the children were assessed with the Wechsler Intelligence Scale for Children – Revised (WISC-R, Wechsler, 1976) and the Neale Reading Comprehension Test (Reynell & Gruber, 1990, p. 56).

- RDLS Verbal Comprehension scores at age 5 showed correlations of .74, .76, and .66 with age 10 WISC Full Scale, Verbal, and Performance IQ, respectively, and a correlation of .67 with scores on the Neale Reading Comprehension Test.
- RDLS Expressive Language scores at age 5 showed correlations .75, .76, and .70 with age 10 WISC-R Full Scale, Verbal, and Performance scales, respectively, and a correlation of .59 with Neale Reading Comprehension Test scores.

Demographic Characteristics and RDLS Scores
The authors report a set of analyses examining differences in RDLS scores based on gender, race/ethnicity, parent education, parent occupation, and geographic region in the U. S. standardization sample. All main effects for these 5 demographic characteristics were significant for Verbal Comprehension scales and all but one main effect was significant for the Expressive Language scale (Reynell & Gruber, 1990, p. 57).

- *Parent education*. Standard scores were lowest for children whose parents had less than high school graduation, followed in order by children whose parents had graduated from high school, those whose parents had some college education, and those whose parents had a 4-year college degree or higher education.
- *Parent occupation*. Standard scores were lowest for children whose parents held unskilled or semiskilled jobs, followed in order by children of parents with skilled or service positions, children of parents holding administrative or lower professional positions, and children of parents with executive or higher professional occupations.
- *Geographic region*. Regional differences were not entirely consistent across RDLS scales. However, children living in East and North–Central regions of the U. S. had higher scores on all RDLS scales than did children living in the South and West.
- *Sex*. Small but significant sex differences were also noted, with girls scoring slightly higher on all scales than boys.
- *Ethnicity/race*. Small but significant ethnic group differences were also found. On average, white and Hispanic children's Verbal Comprehension (VCA and VCB) scores were similar and close to the full sample standard score mean of 100, while black children's scores were, on average, 5 to 6 points lower (i.e., mean standard scores of approximately 94 to 95). For Expressive Language, white and black children had average standard scores near 100, while Hispanic children's average standard score was 6 points lower (i.e., 94); however, the ethnicity effect on Expressive Language was not significant. According to the authors, "Differences associated with ethnicity…tended to be large enough to warrant clinical and interpretive interest…" (p. 58).

Reliability/Validity Information from Other Studies
The following clinical studies were cited in the manual as other studies that address the validity of RDLS scales (Reynell & Gruber, 1990, p. 53).

- Howell, Skinner, Gray, and Broomefield (1981) found that RDLS scores differentiated children referred for speech therapy from a matched comparison group.
- RDLS Expressive Language scale scores were found to be related to executive syntactic competence in children with and without language disorders. In the same study, RDLS Verbal Comprehension scores were related to syntactic speech in a normal language group, but not in a sample of children with language disorders (Udwin & Yule, 1982).

- Studying a sample of autistic children, RDLS Verbal Comprehension and Expressive Language scales were found to be related to measures of spontaneous utterances and grammatical competence (Cantwell, Howlin, & Rutter, 1977).

Comments

- Overall, internal consistency reliability of the RDLS appears to be good, although reliability estimates tended to drop with age, particularly after age 5. As indicated by the authors, results for these older children should be interpreted with caution. The authors further note, however, that "…the RDLS does provide means for dealing with this issue. For example, developmental interpretation is only supported up to age 5 years 0 months. Similarly, for standard score interpretation, all RDLS interpretive materials explicitly and clearly recommend the use of confidence intervals that take scale reliability into account" (Reynell & Gruber, 1990, p. 50).
- The authors give a substantial justification for the manner in which items and sections fit the underlying constructs, but there are very few citations, and it is not clear how the underlying developmental perspective on language development agrees with current thinking on language development. The measure's authors do note that despite the justification for developmental order of the items, children may miss questions considered to be developmentally less challenging, while accurately responding to more developmentally challenging questions.
- Although not discussed in the construct validity section of the manual, results of analyses in which the Rasch model (Rasch, 1980) was applied to data from the U. S. standardization sample do provide some support for the hypothesized developmental sequence of items on the RDLS. As described by the authors, this model "…produces a single continuous scale that can be used to measure two important characteristics over the full range of a test's discrimination: the respondent's ability and the difficulty of each item…" (Reynell & Gruber, 1990, p. 58). Ability/Difficulty scales for Verbal Comprehension (VCA and VCB) and Expressive Language scales were created using the Rasch model. In general, items on the three Ability/Difficulty scales are arranged according in the proposed developmental order (see test booklet, p. 8, included at the end of the manual).
- Split-half reliability, which as noted above was quite high although it tended to drop for children ages 5 years and older, was cited as evidence of construct validity. The authors did not address the extent to which Expressive Language and Verbal Comprehension truly constituted separable language constructs. In the New Zealand study used to examine concurrent validity, the reported correlation between Verbal Comprehension and Expressive Language scores was .76. This finding may suggest that expressive and receptive language as assessed by the RDLS may not reflect distinct language skills, but rather different aspects of a single underlying construct.
- Concurrent correlations with the various criterion measures were generally in the moderate to strong range. As predicted by the authors, somewhat higher correlations were typically found with other measures of verbal and linguistic skills, while correlations with measures involving performance, visual, and manual skills were somewhat lower.
- Predictive validity reported in the manual indicated generally strong correlations between both RDLS scales and well-established measures of IQ. This was noted for two samples,

the New Zealand sample, with assessments conducted at a 3-year time interval, and a British sample in which assessments were made 5 years apart. It is not clear why longitudinal analyses were not conducted (or reported) across ages 4 and 5 for the New Zealand sample. Such analyses could have paralleled the reported concurrent validity analyses.

- It is worth noting that the sample in the British study used for predictive validity analyses included only clinically diagnosed language delayed children, and half of the children in the New Zealand sample had experienced neonatal problems. The extent to which these sample characteristics affected predictive relationships among the measures, and the extent to which associations of similar strength would be found in more normative samples, is not known.
- It should also be noted that the concurrent and predictive validity analyses described in the manual were conducted with data from non-U. S. samples, and involved the British version of the RDLS. The authors note that differences between the two versions are subtle, but it does not appear that concurrent and predictive validity information is available that is specific to the U. S. version.

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

- The RDLS has been found to be quite sensitive to environmental variation in studies of child care. Characteristics of child care settings that are subject to regulation (e.g. caregiver education, training, recent training, number of children enrolled, license status), as well as characteristics that cannot be as directly regulated (e.g. caregiver experience, caregiver age, caregiver mental health, presence of a caregiver's own child) have been found to be related to RDLS outcomes (Clarke-Stewart, Vandell, Burchinal, O'Brien, & McCartney, 2002).
- Various aspects of socioeconomic status, such as income-to-needs levels and parental education level have been found to associate with language development using this measure (Clarke-Stewart et al 2000; NICHD Early Childcare Research Network, 1999, 2000, 2001; Dearing, McCartney, & Taylor, 2001).
- Maternal sensitivity and depression, child gender, and child attachment security have all been found to predict RDLS outcomes (NICHD Early Childcare Research Network, 1999, 2000).

V. Adaptations of Measure

The U. S. version of the RDLS is itself an adaptation of the original British version.

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 the authors

"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, *et al.*, 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 were 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 SICD-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, *et al.*, 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, *et al.*, p. 45).

Validity Information in the Manual

Construct/Concurrent Validity

- Reported correlations of SICD-R Receptive Communication Age (RCA) and Expressive Communication Age (ECA) scores with each other and with and scores on other measures of language development ranged from .74 to .95. Although intercorrelations among scales of the same measure are generally considered as evidence of construct validity, and correlations with separate measures as concurrent validity, this designation was not made by the authors. All reported correlations were high, with the strongest correlation being between the two SICD-R scores (Hedrick, *et al*., 1984, p. 46).
    - RCA and ECA: $r = .95$.
    - RCA and Peabody Picture Vocabulary Test (PPVT; Dunn, 1965): $r = .81$.
    - ECA and PPVT: $r = .76$.
    - ECA and Mean Length of Response: $r = .76$.
    - ECA 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.
- All reported reliability estimates were high.
- More detailed information on the validity of the SICD-R would be useful. However, the information that is provided by Hedrick, *et al.* (1984) suggests that this measure demonstrates high correlations with other language measures with which it should be expected to correlate.

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 SICD-R scores at 12, 24 and 36 months, net of child and family factors (Burchinal, Roberts, Riggins, Zeisel, Neebe & 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.
- It would be useful to have further predictive and construct validity information, 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.

Studies Using Adaptation
O'Reilly, R. (1981). *Language testing with children considered difficult to test* (Masters Thesis). Arizona State University.

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.

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 included 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. The child may be asked to point to a picture illustrating a particular word or to point to a body part. The child may be asked to use a toy in making his or her response. 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 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, how many sentences the child uses to answer open-ended questions, and whether the child uses pronouns properly. At the most difficult level, the child is asked questions that require greater detail to answer (e.g., "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 all 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, *et al*., 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, *et al*., 1999, p. 85). For the Receptive and Expressive subtests, test-retest correlations were highest for the youngest children (ages 2 to 4) and slightly lower (although still high) 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, *et al*., 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, 2000a); 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, 2000b): 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 selected intelligence test scores and TELD-3 scores (Forms A and B, respectively) are presented below (Hresko, *et al*., 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.

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.

Comments
- TELD-3 test-retest reliability was high at all ages, although correlations were slightly higher for younger children.
- 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.
- Although 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.

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

None found.

V. Adaptations of Measure

None found.

References for Language Measures

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

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

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

Brownell, R. (2000c). *Expressive One-Word Picture Vocabulary Test–Spanish-Bilingual Edition*. 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*, 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,* 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.

Cantwell, D., Howlin, P., & Rutter, M. (1977). The analysis of language level and language function: A methodological study. *British Journal of Disorders of Communication, 12*, 119-135.

Carrow-Woolfolk, E. (1985). *Test for Auditory Comprehension of Language- Revised Edition.* Austin, TX: PRO-ED.

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

Clarke-Stewart, K. A., Vandell, D. L., Burchinal, M. R., O'Brien, M., & McCartney, K. (2002). Do features of child care homes affect children's development? *Early Childhood Research Quarterly, 17*, 52-86.

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

Dearing, E., McCartney, K., & Taylor, B. A. (2001). Change in family income-to-needs matters more for children with less. *Child Development, 72*, 1779-1794.

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 Service.

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

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,* 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*, 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.

Frankenburg, W. K. & Bresnick B. (1998). *Denver II Prescreening Questionnaire*. Denver, CO: Denver Developmental Materials.

French, J. L. (1964). *Pictorial Test of Intelligence*. Boston, MA: Houghton Mifflin.

Frostig, M. (1966). *Developmental Test of Visual Perception*. Palo Alto, CA: Consulting Psychologists Press.

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.

Goodenough, G. L., & Harris, D. B. (1963). *Goodenough-Harris Drawing Test*. San Antonio, TX: The Psychological Corp.

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

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

Howell, J., Skinner, C., Gray, M., & Broomefield, S. (1981). A study of the comparative effectiveness of different language tests with two groups of children. *British Journal of Disorders of Communication, 16*, 31-42.

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.

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.

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

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.

Kirk, S. A., McCarthy, J. J., & Kirk, W. D. (1968). *The Illinois Test of Psycholiguistic Abilities-Revised*. Champlain, IL: University of Illinois Press.

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,* 270-275.

Locke, A., Ginsborg, J., & Peers, I. (2002). Development and Disadvantage: Implications for the early years and beyond. *International Journal of Languages and Communication Disorders, 37*, 3-15.

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*, 244-260.

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

Metropolitan Achievement Test, Seventh Edition. (1992). San Antonio, TX: Harcourt Brace Educational Measurement.

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

Mott, S. E. (1987). Concurrent validity of the Battelle Developmental Inventory for speech and language disordered children. *Psychology in the Schools, 24*, 215-220.

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

Newcomer, P. & Hammill, D. (1997). *Test of Language Development- Primary.* Austin, TX: PRO-ED.

NICHD Early Child Care Research Network. (1999). Chronicity of maternal depressive symptoms, maternal sensitivity, and child functioning at 36 months. *Developmental Psychology, 35*, 1297-1310.

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

NICHD Early Child Care Research Network. (2001). Child-care and family predictors of preschool attachment and stability from infancy. *Developmental Psychology, 37*, 847-862.

NICHD Early Child Care Research Network (2002a). Early child care and children's development prior to school entry: Results from the NICHD Study of Early Child Care. *American Educational Research Journal*, *39*, 133-164.

NICHD Early Child Care Research Network (2002b). Child-care structure ← process ← outcome: Direct and indirect effects of child-care quality on young children's development. *Psychological Science, 13*, 199-206.

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

Otis, A. & Lennon, R. (1995). *Otis-Lennon School Ability Test, Seventh Edition.* 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,* 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*, 451-477.

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. (Original work published in Copenhagen, 1960).

Reynell, J., & Gruber, C. P. (1990). *Reynell Developmental Language Scales - U.S. Edition.* Los Angeles: Western Psychological Services.

Semel, E., Wiig, E., & Secord, W. (1995). *Clinical Evaluation of Language Fundamentals, Third Edition*. San Antonio, TX: The Psychological Corp.

Stanford Achievement Test, Ninth Edition. (1996). San Antonio, TX: Harcourt Brace Educational Measurement.

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

Terman, L. M., & Merrill, M. R. (1960). *Stanford-Binet Intelligence Scale*. Boston: Houghton Mifflin.

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,* 347-353.

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

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

Udwin, O., & Yule, W. (1982). A comparison of performance on the Reynell Developmental Language Scales with the results of a syntactical analysis of speech samples. *Child: Care, Health and Development, 8*, 337-343.

Van Riper, C., & Erickson, R. L. (1968). *Predictive Screening Test of Articulation.* Kalamazoo, MI: Western Michigan University, Continuing Education Office.

Vernon-Feagans, L., Emanuel, I. B. & Blood, I. (1997). The effect of otitis media and quality daycare on children's language development. *Journal of Applied Developmental Psychology*, 18, 395-409.

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

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

Wechsler, D. (1976). *Wechsler Intelligence Scale for Children-Revised*. New York: The Psychological Corp.

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

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children, Third Edition.* San Antonio, TX: The Psychological Corp.

Wiig, E., Secord, W., & Semel, E. (1992). *CELF- Preschool: The Clinical Evaluation of Language Fundamentals - Preschool Examiner's Manual*. San Antonio, TX: The Psychological Corp.

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. L., Steiner, V. G., & Pond, R.E. (2002). *Preschool Language Scale Fourth Edition: Examiner's Manual*. San Antonio, TX: The Psychological Corp.

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

Literacy Measures

Dynamic Indicators of Basic Early Literacy Skills 6th Edition (DIBELS)

I. Background Information

Author/Source

Source: Good, R. H. & Kaminski, R. A. (Eds.). (2002a). Dynamic Indicators of Basic Early Literacy Skills (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Publisher: Institute for the Development of Educational Achievement
1211 University of Oregon
Eugene, Oregon 97403-1211
Website: http://dibels.uoregon.edu

Purpose of Measure

As described by the authors

The purpose of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) is to assess the early literacy domains discussed in reports from the National Reading Panel (2000) and the National Research Council (1998). Five measures included in the DIBELS were designed to assess 3 of 5 "foundational skills related to reading outcomes," or "big ideas" of early literacy (Good, Gruba, & Kaminski, 2002, p. 682). "[Big ideas] are skills that differentiate successful from less successful readers and, most important, are amenable to change through instruction…" (Good *et al.*, p. 682). The three foundational skills that are the focus of DIBELS measures include Phonological Awareness, Alphabetic Principle, and Fluency with Connected Text.

DIBELS measures "…are designed to be short (one minute) fluency measures used to regularly monitor the development of pre-reading and early reading skills. …to assess student development of phonological awareness, alphabetic understanding, and automaticity and fluency with the code. …When used as recommended, the results can be used to evaluate individual student development as well as provide grade-level feedback toward validated instructional objectives" (http://dibels.uoregon.edu/data/DIBELS_Data_System_Desc.pdf, p. 1).

Population Measure Developed With

There is currently no published unified source for information on the samples used in the development of all DIBELS measures. A published report including Phonemic Segmentation and Letter Naming Fluency measures (Kaminski & Good, 1996) describes a sample of 37 kindergarten and 41 first grade students from a predominantly white, rural elementary school in the Pacific Northwest. About 9 percent of the population received special education and 12 percent were eligible for the free lunch/milk program (p. 4).

Percentile scores have been developed using data from all schools that use the DIBELS system. This constitutes a national, but not necessarily nationally representative, sample including slightly less than 40,000 kindergartners, 40,000 first graders, approximately 15,000 second graders, and more than 10,000 third graders.

Age Range Intended For

Preschool through third grade. The specific measures that are recommended vary by age.

Key Constructs of Measure

The DIBELS consists of six measures—Initial Sounds Fluency, Phonemic Segmentation Fluency, Nonsense Word Fluency, Oral Reading Fluency and Retell Fluency, Letter Naming Fluency, and Word Use Fluency. Children's scores on each of these measures can be used in several ways. Scores can be evaluated based on normative information – local school district norms as well as norms established by the authors from information provided by participating users of the DIBELS Web data system. In addition, the authors have developed benchmarks for individual measures that reflect "…goals for the lowest-achieving student in the school…DIBELS benchmark goals are the minimal level students need to achieve to be confident they are on track for literacy outcomes…" (Good, Gruba, & Kaminski., 2002, p. 684). Finally, the authors provide descriptive levels (At Risk/Deficit, Some Risk/Emerging, and Low Risk/Established) for individual measures, and instructional support recommendations (Benchmark – At Grade Level, Strategic – Additional Intervention, and Intensive – Needs Substantial Intervention) based on patterns of performance across measures. These recommendations are based on the percentages of children demonstrating each possible pattern of performance who achieve subsequent literacy goals (Good, Simmons, Kame'enui, Kaminski, & Wallin, 2002).

- *Initial Sound Fluency*. This measure is intended for children from the beginning of the last year of preschool through the middle of kindergarten; a benchmark has been established for this measure in the middle of kindergarten. It is used to assess the ability of a child to produce and recognize the first sound of an orally presented word (Good, Laimon, Kaminski & Smith, 2002, p. 10).
- *Letter Naming Fluency*. This measure is intended for most students from the beginning of kindergarten through the fall of first grade. The test provides a measure of how freely children label letters of the alphabet. This construct is not considered essential to reading achievement and no benchmark has been established, but children whose percentile scores based on local norms reflect relatively poor performance may be at risk for negative reading outcomes (Kaminski & Good, 2002, p. 6).
- *Word Use Fluency*. This measure is intended for children from the beginning of kindergarten through the end of third grade; no benchmarks have been established because "…additional research is needed to establish its linkage to other big ideas of early literacy…" (Good, Kaminski, & Smith, 2002a, p. 39). This measure taps the accuracy and complexity of how children use words.
- *Phoneme Segmentation Fluency*. This measure is intended for children from the middle of kindergarten through the end of first grade; benchmark goals have been established for the spring of kindergarten and for the fall of first grade. It is used to assess the student's ability to segment three and four phoneme words (Good, Kaminski & Smith, 2002b, p.16).
- *Nonsense Word Fluency*. This measure is intended for children from the middle of kindergarten through the beginning of second grade; a benchmark goal has been established for the middle of first grade. It tests the alphabetic principle of letter-sound correspondence, as well as the ability to blend letters into words (Good & Kaminski, 2002b, p. 23).

- *Oral Reading Fluency and Retell Fluency*. These measures are intended for children from the middle of first grade through the end of third grade. They are based on Curriculum-Based Measurement (CBM) procedures and measure the fluency and accuracy with which children read short passage of text (Good, Kaminski & Dill, 2002, p. 30). Benchmarks have been established for the spring of first grade, the spring of second grade, and the spring of third grade. Retell Fluency is primarily used to "…provide a comprehension check for the [Oral Reading Fluency] assessment" (p. 30), and no separate benchmarks have been established.

Norming of Measure (Criterion or Norm Referenced)
Both Norm and Criterion referenced. Norms (percentile rankings) are based on the population of students whose scores have been entered into the DIBELS Web data system by participating schools and districts (i.e., they are not based on a nationally-representative random sample). The use of local norms is also recommended for interpreting children's scores.

Comments
Information on the samples of children used in developing DIBELS measures is not summarized in the chapters that comprise the DIBELS manual, and such information is not consistently available from published reports. In order to evaluate fully the usefulness of DIBELS measures, it would be useful to have such a unified resource for users and potential users.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
The tasks that children are asked to perform vary by the measure being administered. For Letter Naming, the child is given one minute to identify orally as many randomly placed upper and lower case letters on a page as he or she can. For Initial Sound Fluency, children are asked to identify the beginning sounds of pictured objects. The examiner names the four pictures, then says the beginning sound of each picture. The child must point to the picture that begins with the sound. In the second section of this measure, children are also required to produce the beginning sound. For Phoneme Segmentation, children are asked to pronounce the individual phonemes that make up each word the examiner presents verbally. For Nonsense Word Fluency, children are presented a paper with a series of nonsense words (e.g., ov, vaj, sig) printed on it and are asked to pronounce the individual phonemes or to read the nonsense word. For Oral Reading Fluency, children are given passages to read aloud. The Retell Fluency measure addresses whether children understand the content of the Oral Reading Fluency passage. For Word Use Fluency, children are given a word and are asked to use it in a phrase.

Who Administers Measure/ Training Required?
Test Administration
The authors state that most educational personnel can be trained to administer the measures (http://dibels.uoregon.edu/dibels_how.php).

Data Interpretation

- The authors state that most educational personnel can be trained to score DIBELS measures. Interpretation is relatively straightforward and is based on local norms as well as percentile scores and benchmarks provided by the authors (http://dibels.uoregon.edu/dibels_how.php).
- The authors provide training for the administration and interpretation of DIBELS.

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

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

Time Needed and Cost

Time

Each test is designed to take from one minute to three minutes, and the entire assessment should take less than 10 minutes per child.

Cost

- There is no cost at this time. Materials can be downloaded from the website and copies can be made.
- Sopris West publishes a printed version of the measures for $59—enough for a classroom of 25 students. The materials can be ordered online at www.sopriswest.com.
- School districts can also enter their data online to receive automated reports. The web-based service costs $1 per student per year.

III. Functioning of Measure

Reliability Information from Manual

The DIBELS Administration and Scoring Guide (Good & Kaminsky, 2002a) includes summary information on the reliability of DIBELS measures. Each measure is described in a separately authored chapter. Most of the reliability information involves alternate form reliability; because DIBELS measures are designed to be used for ongoing assessment, each measure has numerous alternate forms.

- The Initial Sound Fluency measure is a minimally revised version of another measure, Onset Recognition Fluency. Reliability information is based on studies using the Onset Recognition Fluency measure. The reported alternate-form reliability of that measure was .72 in January of the kindergarten year. When averaged across 4 assessments, reliability increased to .91 (Good, Laimon, Kaminski & Smith, 2002, p. 10).
- Letter Naming Fluency had a one-month, alternate-form reliability of .88 in kindergarten (Kaminski & Good, 2002, p. 6).
- Reliability information was not available for Word Use Fluency.
- Phoneme Segmentation Fluency showed a two-week alternate-form reliability of .88, and a one-month alternate-form reliability of .79 at the end of kindergarten (Good, Kaminski & Smith, 2002b, p. 16). For Nonsense Word Fluency, the reported one-month alternate-form reliability was .83 for the middle of first grade (Good & Kaminski, 2002b, p. 23).

- No specific reliability information was provided for Oral Reading Fluency and Retell Fluency. These measures are based on Curriculum-Based Measurement (CBM) procedures for reading developed by Deno and colleagues (e.g., Tindal, Marston, & Deno, 1983) and by Shinn (1989). The authors indicated that reliabilities of CBM reading procedures generally range from .92 to .97 and alternate-form reliabilities range from .89 to .94 (Tindal, et al., 1983, as cited in Good, Kaminski & Dill, 2002, p. 30). Time intervals between assessments were not reported.

Validity Information from Manual

Concurrent Criterion-Related Validity

- Correlations were reported between Initial Sound Fluency (Onset Recognition Fluency) scores and two additional measures: 1) DIBELS Phoneme Segmentation Fluency scores and 2) Readiness Cluster standard scores from the Woodcock-Johnson Psycho-Educational Battery-Revised (WJ-R; Woodcock & Johnson, 1989). These correlations were .48, and .36, respectively (Good, Laimon, Kaminski & Smith, 2002, p. 10).
- For Letter Naming Fluency, Kaminski and Good (2002, p. 6) reported that the "median criterion-related validity" of Letter Naming Fluency with WJ-R Readiness Cluster standard scores was .70 in kindergarten.
- No information was provided in the Administration and Scoring Guide regarding the validity of the Word Use Fluency measure.
- The correlation between Phoneme Segmentation Fluency and WJ-R Readiness Cluster standard scores was .54 in the spring of kindergarten (Good, Kaminski & Smith, 2002b, p. 16). Correlations between Nonsense Word Fluency and the WJ-R Readiness Cluster scores in January and February of first grade were .36 and .59, respectively (Good & Kaminski, 2002, p. 23b).
- According to the authors, Oral Reading Fluency showed correlations with criterion measures that ranged from .52 to .91 "in eight separate studies from the 1980s" (Good, Kaminski & Dill, 2002, p. 30). No additional information regarding what criterion measures were used, length of delay between testings, or information about the samples was provided in the manual. The Retell Fluency scale showed a concurrent correlation of .59 with Oral Reading Fluency.

Predictive Validity

- Correlations between Initial Sound Fluency assessed in kindergarten and criterion measures assessed a year later in the children's first-grade year were .45 with DIBELS Oral Reading Fluency scale and .36 with the WJ-R Reading Cluster score (Good, Laimon, Kaminski & Smith, 2002, p. 10).
- Correlations between Letter Naming Fluency assessed in kindergarten and criterion measures assessed a year later in the children's first-grade year were .65 with WJ-R Reading Cluster scores in kindergarten and .71 with scores from a first grade CBM procedure (Kaminski & Good, 2002, p. 6).
- Correlations between kindergarten Phoneme Segmentation Fluency and criterion measures in the children's first-grade year were .62, .62 and .68 with DIBELS Nonsense Word Fluency, CBM oral reading fluency, and WJ-R Reading Cluster scores, respectively (Good, Kaminski & Smith, 2002b, p. 16).

- Correlations between Nonsense Word Fluency in January of first grade and criterion measures either later that year or the following school year were .82, .60, and .66 with CBM reading procedure scores in May of first grade, and CBM reading procedure and WJ-R Reading Cluster scores in May of second grade, respectively (Good & Kaminski, 2002b, p. 23).
- No information is provided in the DIBELS Administration and Scoring Guide regarding the predictive validity of the Oral Reading Fluency, Retell Fluency and Word Use Fluency measures.

Reliability/Validity Information from Other Studies

Kaminski and Good (1996) examined the reliability and validity of three DIBELS measures, Phoneme Segmentation Fluency, Letter Naming Fluency and Picture Naming Fluency (no longer included in DIBELS) with two cohorts—one of kindergartners (n = 37) and one of first graders (n = 41). Half of the children in each age group were tested twice, once at the beginning and once at the end of a nine-week period, while the remaining children were tested with alternate forms of DIBELS measures twice each week over the same nine week period.

- Alternate form reliabilities were computed for the half of the sample that was tested twice weekly with DIBELS measures. All possible pairs of assessments that were separated by one week or less were included in these analyses. For kindergartners, reliabilities were .93 for Letter Naming and .88 for Phoneme Segmentation. For first graders, reliabilities were .83 for Letter Naming and .60 for Phoneme Segmentation.
- An additional method used to examine reliability was to average odd and even numbered assessments separately, and to correlate these averaged scores. For kindergartners, correlations were .99 for both Letter Naming and Phoneme Segmentation. For first graders, correlations were .95 for Letter Naming and .83 for Phoneme Segmentation.
- The concurrent validity of DIBELS measures was examined by correlating DIBELS scores with scores on a number of criterion measures. These included the McCarthy Scales of Children's Abilities (McCarthy, 1972), the Rhode Island Pupil Identification Scale (Novack, Bonaventura, & Merenda, 1973), a teacher rating scale covering reading readiness/achievement, rate of progress in reading readiness/achievement, and level of risk for later reading problems, the Metropolitan Readiness Test, Level 2 (Nurss & McGauvran) administered to kindergartners only, the Stanford Diagnostic Reading Test (Karlen & Gardner, 1985) administered to first graders only, and CBM reading procedures (Marston, 1989), also administered only to first graders.
    - For kindergarten children, correlations between Letter Naming and the McCarthy Scales of Children's Abilities, the Metropolitan Readiness Test, the Rhode Island Pupil Identification Scale, and the teacher rating scale were .59, .77, .67, and .85 respectively.
    - First grade correlations between Letter Naming and the McCarthy Scales, the Stanford Diagnostic Reading Test, the Rhode Island Pupil Identification Scale, the teacher rating scale, and CBM reading procedure scores were .13, .50, .27, .34, and .45, respectively.
    - Correlations between Phoneme Segmentation and the McCarthy Scales of Children's Abilities, the Metropolitan Readiness Test, the Rhode Island Pupil Identification Scale, and the teacher rating scale were .46, .65, .43, and .55, respectively, for kindergarten children.

- First grade correlations between Phoneme Segmentation and the McCarthy Scales of Children's Abilities, the Stanford Diagnostic Reading Test, the Rhode Island Pupil Identification Scale, the teacher rating scale, and the CBM reading measure were .02, .29, .06, .09, and .09, respectively.

Good and colleagues (Good, Simmons, & Kame'enui, 2001) examined the utility and predictive validity of the DIBELS with four cohorts of students from kindergarten through third grade. The authors assessed the utility of the DIBELS benchmarks by analyzing whether they predicted reading fluency over time, as well as whether they predicted outcomes on a high stakes statewide reading/literature assessment. They found that 96 percent of students who met the third grade oral reading fluency benchmark goal also met or exceeded expectations for the statewide assessment.

Hintze and colleagues (Hintze, Ryan, & Stoner, 2003) examined the concurrent validity of DIBELS Letter Naming Fluency, Initial Sound Fluency, and Phoneme Segmentation Fluency, by correlating these measures with measures from the Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgesen, Rashotte, 1999). The sample consisted of 86, primarily white (93%) kindergartners. The correlations ranged from .08 to .63, with a median of .46. Using Receiver Operator Characteristic curve modeling, DIBELS benchmarks were shown to have high levels of sensitivity in identifying children at risk for future reading problems when compared to CTOPP criterion; at the same time, DIBELS benchmarks illustrated low specificity (i.e., resulted in many false positives). The authors presented a set of new suggested benchmarks, which were shown to improve DIBELS specificity.

Comments

- Reliability information that is readily available for the DIBELS shows strong alternate-form reliability across the measures. This is particularly important given that these measures are intended to be used with children repeatedly over time. Other than the test-retest reliability noted for the Oral Reading Fluency and Retell Fluency, however, very little other information about reliability is available. For instance, no information is provided regarding interrater reliability. Information regarding interrater reliability may be of particular importance when the assessors are classroom teachers, rather than assessment specialists.
- When validity information is available in the Administration and Scoring Guide, it generally indicates moderate to strong concurrent and predictive validity. The details of how validity was assessed for Oral Reading Fluency are not provided in the manual. It is simply stated that there were strong correlations between this scale and criterion measures in "eight separate studies in the 1980s" (Good, Kaminski & Dill, 2002, p. 30). There is no psychometric information available on Word Use Fluency from the Administration and Scoring Guide, and we were unable to find any other published source for such information.
- Validity information from published studies reported widely ranging correlations with criterion variables. Of particular note, Kaminski and Good (1996) reported moderate to high correlations with criterion measures for kindergartners, but only low to moderate correlations with criterion variables for first graders (with the exception of a .50 correlation between Letter Naming Fluency and Stanford Diagnostic Reading Test

scores). The authors indicated that these DIBELS measures might have been too easy for first graders, resulting in ceiling effects that could help to explain the low correlations at the older age.

- Overall, the available reliability and validity information for DIBELS measures is limited, and some of the available information is based on earlier versions of the measures. Little information is currently provided in the manual or in published sources regarding the samples that were used to determine the reliability and validity of the measures. The available information does not allow an evaluation of whether reliability and validity may systematically differ for groups of children who differ on important characteristics (e.g., ethnicity, SES, gender). Further, given current information it is difficult to assess whether psychometric characteristics of the measures differ by child age. As noted above, early findings reported in Kaminski and Good (1996) suggest that both reliability and validity were consistently higher in the kindergarten sample than in the first grade sample. In order to determine the usefulness of these measures across the age range for which they are intended, it will be important for information regarding samples used for DIBELS development and evaluations of psychometric properties to be provided in much greater detail. Drawing together the psychometric information, with adequate description of samples and measures used, into a single source (i.e., a manual) would be a valuable addition to researchers and practitioners desiring to use DIBELS measures.

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

Kamps, Wills, Greenwood, Thorne, Lazo, Crockett, Akers, and Swaggart (2003) used DIBELS Letter Naming Fluency, Nonsense Word Fluency, and Oral Reading Fluency as measures of academic risk and as outcome measures in a study comparing the effectiveness of three reading programs—a literature-based program, Reading Mastery (1995), and Success for All (1999). The two-cohort, ethnically diverse sample used in this study included 383 children (213 boys) attending five different schools. Children were originally assessed in kindergarten, first, or second grade, and were followed for two or three years (depending upon initial grade and cohort). Approximately half (197, or 51.4 percent) of the children were initially identified as being at academic risk, based on DIBELS scores; 60 of these children were identified as having behavioral risk factors as well. Forty additional children had behavioral risk factors but were not identified as being at academic risk. The authors interpreted their findings to suggest that, while all children demonstrated increases in reading skills as reflected in DIBELS scores, the Reading Mastery curriculum promoted the largest gains, followed by Success for All. The literature-based program resulted in the lowest gains. Further, the results indicated that children with behavioral or academic risks demonstrated lower gains than did children without either type of risk alone. Children with behavioral risk only did better than children with academic risk, and children with both behavioral and academic risk demonstrated the smallest gain on DIBELS measures across time.

Gunn, Smolkowski, Biglan, and Black (2002) examined the effects of supplemental reading instruction intervention for students in kindergarten through third grade. The intervention was

presented over two academic years; effects were assessed one year following the end of the intervention. The sample included 256 students (142 boys) who were identified as being below grade level in reading, 100 of whom were also identified as exhibiting aggressive social behavior. Approximately 62 percent of the children were Hispanic (94 percent of Mexican heritage, 85 percent born in Mexico, 84 percent exclusively or nearly exclusively Spanish speakers). Several DIBELS measures were included to screen children for reading difficulties. An Oral Reading Fluency measure similar to that included in the DIBELS was used as an outcome measure of reading achievement. Children were randomly assigned to receive the supplemental reading instruction intervention, or to a control group that did not receive supplemental instruction. The supplemental instruction consisted of four to five months of instruction during the first year of the study, and nine months during the second year. Trained instructional assistants provided all supplemental instruction. Results indicated that both Hispanic and non-Hispanic children in the intervention group demonstrated significantly greater gains on the Oral Reading Fluency measure than did children in the control group.

V. Adaptations of Measure

DIBELS-M

Elliot and colleagues (Elliott, Lee, & Tollefson, 2001) examined the psychometric properties of DIBELS Letter Naming Fluency and Sound Naming Fluency, and two other slightly modified DIBELS scales, Initial Phoneme Ability (adapted from Initial Sound Fluency), and Phonemic Segmentation Ability (adapted from Phonemic Segmentation Fluency). The two modified scales differed from the DIBELS only in that they used stimulus words that were slightly less difficult, and the assessments were not timed. Collectively, this version was called the DIBELS-M. A sample of 75 kindergarten students were administered the four scales twice at the end of their school year, with a two-week interval between administrations. In addition to the completion of the DIBELS-M, the kindergartners also completed the WJ-R Broad Reading and Skills clusters, the Test of Phonological Awareness—kindergarten form (TOPA; Torgesen & Bryant, 1994), and the Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990). The researchers also used data from the Developing Skills Checklist (DSC; 1990) that was administered at the beginning of the year and an informal teacher-rating questionnaire. Elliot et al (2001) reported three types of reliability estimates for the DIBELS-M (interrater, test-retest and alternate forms). Interrater reliabilities ranged from .82 to .94, test-retest reliabilities ranged from .74 to .93, and alternate form reliabilities ranged from .64 to .91. The DIBELS-M had the highest concurrent correlations with the WJ-R Skills Cluster and the DSC. Correlations ranged from .60 to .81 with the WJ-R Skills Cluster and ranged from .54 to .74 with the DSC. These findings are consistent with earlier findings noting a significant concurrent relationship between the WJ-R and the DIBELS.

Spanish Language Version

A Spanish version of the DIBELS is available.

Test of Early Reading Abililty-3 (TERA-3)

I. Background Information

Author/Source

Source: Reid, D. K., Hresko, W. P., & Hammill, D. D. (2001). Test of Early Reading Ability-3. Austin, TX: PRO-ED, Inc.

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

Purpose of Measure

As described by the authors

This measure assesses emerging reading skills in young children. "The TERA-3 has five purposes: (a) to identify those children who are significantly below their peers in reading development and thus may be candidates for early intervention, (b) to identify strengths and weaknesses of individual children, (c) to document children's progress as a consequence of early reading intervention programs, (d) to serve as a measure in research studying reading development in young children, and (e) to accompany other assessment techniques" (Reid, Hresko & Hammill, 2001, p.8).

The authors caution that this assessment does not measure children's reading ability at a fine enough level to address specific student problems or inform individual instruction. The authors encourage that the assessment of young children's reading ability be based on a variety of sources of information (e.g., direct observation, parent questionnaires, teacher questionnaires).

Population Measure Developed With

The norming sample for the TERA-3 included 875 children from 22 U.S. states, ranging in age from 3 years, 6 months to 8 years, 6 months. The sample included children attending day care centers, preschools and schools based in public, private and church facilities. In addition, professionals who had used the previous version of the TERA (TERA-2; Reid, Hresko, & Hammill, 1989), who had participated in developing other tests for the publisher, or who were colleagues of the authors were also recruited to provide data on children in their area. A nationally representative sample was obtained with respect to geographic region (Northeast, Midwest, South, and West), gender, race (white, black, or other), ethnicity (African American, Hispanic American, Asian/Pacific Islander, Native American/Eskimo/Aleut, and European American/Other), urban versus rural residence, family income, educational attainment of parents (less than a bachelor's degree, bachelor's degree, or advanced degrees), and disability status (no disability, learning disability, speech-language disorder, mental retardation, or other). With respect to disability status, 5 percent of the sample were diagnosed with a learning disability, 3 percent had an identified speech-language disorder, 1 percent were diagnosed with mental retardation, and 1 percent had some other identified disability. All children completed all items from Form A and Form B, the two alternate forms of the test (Reid *et al.*, 2001, pp. 37-40).

Age Range Intended For
3 years, 6 months to 8 years, 6 months.

Key Constructs of Measure
The TERA-3 consists of three subtests that assess early reading skills. Scores for each subtest are expressed as standard scores with means of 10 and standard deviations of 3. In addition to subtest scores, a Reading Quotient or Composite (both terms are used by the authors) can be derived, which is the standardized (mean = 100, standard deviation = 15) sum of the standard scores from each of the three subtests. The three subtests are:

- *Alphabet*. The Alphabet subtest measures children's knowledge and use of letters. Children who do well on this subtest usually have age-appropriate understanding of phonics and decoding printed words; children who do less well may confuse or mispronounce letters.
- *Conventions*. The Conventions subtest measures children's understanding of print conventions in the English language. Children who do well on this subtest know to read from left to right, understand the uses of punctuation, and know how to hold a book and turn pages; children who do less well may fail to notice punctuation or may not know standard reading conventions.
- *Meaning*. The Meaning subtest measures children's ability to find meaning in print. Children who do well on this subtest can usually recognize the meaning of words in a variety of contexts. In addition, children can retell stories that they have read. Children who do less well have difficulty retelling stories and may often lose their place when reading (Reid *et al.*, 2001, pp. 29-30).

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

Comments

- Standard and percentile scores developed for 3- and 6-month age bands (this varies based on findings from the normative sample) are the scores that are most typically reported for this measure. Age and grade equivalents are also provided. However, the authors state that "Because interpolation, extrapolation, and smoothing were used to create age and grade equivalents, these scores should be interpreted with caution" and further recommend that "…the use of age and grade equivalents [should be used] only on those few occasions when they are mandated by state or federal agencies" (Reid *et al.*, 2001, p. 44).
- Although the sample is described as being nationally representative, it should be noted that the lowest category included for parent education is less than a bachelor's degree. Information on the percentages of parents with a high school diploma or less than a high school education is not provided.
- It is not clear from the manual whether any children from families in which English is not the primary language were included in the sample used for measurement development, and it does not comment on the appropriateness of the use of this test for children from such families. Caution in interpreting results for children with limited exposure to the English language may be warranted.

II. Administration of Measure

Who is the Respondent to the Measure?
Child.

If Child is Respondent, What is Child Asked to Do?
The TERA-3 utilizes entry points, basals and ceilings. What the child is asked to do differs by subtest. The Alphabet subtest requires children to recognize print, name printed letters, identify sounds of words in print, segment words in print, count syllables in print and identify basic sight vocabulary. The Conventions subtest requires children to recognize uppercase and lowercase letters, demonstrate the ability to handle a book, know the function of punctuation and spelling, and differentiate homophones. The Meaning subtest requires the respondent to identify correct use of relational vocabulary, paraphrase, retell a story and identify story topics, use text to predict upcoming events, and understand genre (different uses of text), relate sentences to pictures, and show general awareness of print.

Who Administers Measure/Training Required?
Test Administration
Examiners who administer the TERA-3 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. The authors also suggest that the examiner should have supervised experience in administering standardized tests of mental ability.

Data Interpretation
Interpretation of TERA-3 scores requires more knowledge and experience than that required for administering and scoring the test. Examiners who interpret the TERA-3 scores should have graduate-level training in statistics and in the procedures governing administration, scoring and interpretation of assessments.

Setting (e.g., one-on-one, group, etc.)
The TERA-3 is designed to be administered in a one-on-one setting.

Time Needed and Cost
Time

- The TERA-3 is not a timed assessment. The time required to administer the entire TERA-3 varies from about 15 to 45 minutes. Children below age 5 may need a break approximately every 10 minutes. Children should be given as much time as needed to complete the items. The authors suggest that examiners should rely on their own professional experience to determine a reasonable amount of time for completion of each item and the entire assessment.
- Typically the TERA-3 is completed in one session; however it can be administered over several sessions.

Cost

- TERA-3 Complete Kit (including two equivalent forms and materials): $236
- Manual: $81

III. Functioning of Measure

Reliability Information from Manual

Internal Consistency Reliability

Internal consistency reliabilities (Cronbach's coefficient alphas) were calculated for the subtests and the Reading Quotient (also identified by the authors as the Reading Composite). Coefficient alphas were computed separately for Forms A and B within six age groups. There were 89 3-year-olds, 104 4-year-olds, 160 5-year-olds, 231 6-year-olds, 186 7-year-olds, and 105 8-year-olds.[11] Reliabilities were high for each age group (see Reid *et al.*, 2001, p. 47).

- Alphabet subtest Form A internal reliabilities were .84 at age 3, .94 at age 4, .93 at age 5, .91 at age 6, .93 at age 7, and .88 at age 8. Alphabet subtest Form B internal reliabilities were .83 at age 3, .94 at age 4, .92 at age 5, .92 at age 6, .89 at age 7, and .87 at age 8.
- Conventions subtest Form A internal reliabilities were .81 at age 3, .88 at age 4, .86 at age 5, .80 at age 6, .84 at age 7, and .79 at age 8. Conventions subtest Form B internal reliabilities were .82 at age 3, .89 at age 4, .82 at age 5, .82 at age 6, .87 at age 7, and .75 at age 8.
- Meaning subtest Form A internal reliabilities were .80 at age 3, .94 at age 4, .84 at age 5, .92 at age 6, .91 at age 7, and .91 at age 8. Meaning subtest Form B internal reliabilities were .84 at age 3, .95 at age 4, .84 at age 5, .92 at age 6, .92 at age 7, and .90 at age 8.
- Reading Quotient Form A internal reliabilities were .91 at age 3, .97 at age 4, .95 at age 5, .95 at age 6, .95 at age 7, and .94 at age 8. Reading Quotient Form B internal reliabilities were .91 at age 3, .97 at age 4, .94 at age 5, .95 at age 6, .96 at age 7, and .94 at age 8.
- Coefficient alphas for all subtests and the Reading Composite were also computed separately (across age) for the following subsamples: males, females, European Americans, African Americans, Hispanic Americans, learning disabled children, language impaired children, and reading disabled children. All alphas for all subgroups on both Forms A and B were reported as being .91 or higher (see Reid *et al.*, 2001, p. 48).

As an additional measure if internal consistency, alternate forms reliabilities were estimated by correlating raw scores from Form A with raw scores from Form B. All children in the normative sample received were tested with all items from both forms. Correlations between Forms A and B for the Alphabet subtest ranged from .82 at age 8 to .92 at ages 4 and 6. For Conventions, correlations between Forms A and B ranged from .82 at age 6 to .90 at ages 4 and 8. Correlations between Forms A and B of the Meaning subtest ranged from .88 at ages 6 and 8 to .95 at age 4. No alternate forms reliability estimate was provided form the Reading Composite (see Reid *et al.*, 2001, p. 49).

Test-retest reliability

Test-retest reliabilities were reported for the TERA-3 using two test-retest studies, a Michigan study of 30 children ages 4 to 6 years attending a public school and a Texas study of 34 children ages 7 to 9 years. Children completed each of the subtests on both Forms A and B on two

[11] The sample includes children from 3 years, 6 months through 8 years, 6 months. Although not specified in the manual, it appears that the youngest and oldest age groups in these analyses reflect 6-month age bands, while the 4, 5, 6, and 7-year-old age groups reflect 12-month age bands.

occasions, with a two-week interval between assessments. Test-retest reliability was adequate for the age groups represented in the two studies. For ages 4 to 6 (the Michigan study), test-retest correlations for the combined Forms A and B were .94 for the Alphabet subtest, .87 for the Conventions subtest, .97 for the Meaning subtest, and .98 for the Reading Composite. For ages 7 to 9 (the Texas study), test-retest correlations for the combined Forms A and B were .94 for the Alphabet subtest, .97 for the Conventions subtest, .87 for the Meaning subtest, and .98 for the Reading Composite. Test-retest correlations for the separate Forms A and B were nearly identical to these combined-forms correlations (see Reid *et al.*, 2001, p. 51).

Interrater reliability
Interrater reliability in scoring was reported for the TERA-3 using a randomly selected set of 40 completed protocols from the normative sample. Three raters (one of the authors and two advanced graduate students in special education) participated in independently scoring these protocols. For all three subtests and the Reading Composite, interrater correlations were .99 (see Reid *et al.*, 2001, p. 52).

Validity Information from Manual
Content-Description Validity
The authors discuss in some detail the manner in which content validity (termed content-description validity by the authors, after Anastasi & Urbina, 1997) was addressed during test development. Item development for the TERA-3 involved several approaches.

- First, items from previous versions of the TERA were reviewed and new items were generated following a review of existing research, curricula, and tests. According to the authors, "…we determined that our terminology and content focus reflected that of research…looking at the nature of early reading" and items were selected that also reflected "…current practical thinking in the field" (Reid *et al.*, 2001, p. 57). Further, the authors concluded that "…the content of our subtests is supported by a number of writers…" (p. 59). The authors indicated that an assessment of phonemic awareness was specifically excluded from the TERA-3 because it does not directly involve interacting with print materials.
- Second, items were reviewed by seven professionals with expertise in reading in order to check the authors' placement of items within subtests. For each item, the percentage agreement between subtest placement and the seven experts' opinions regarding appropriate placement was calculated. These percentages were then averaged within subtest. According to Reid *et al.* (2001, p. 62), average agreement for Alphabet subtest items was 99 percent, agreement for Conventions subtest items was 90 percent, and agreement for Meaning subtest items was 98 percent.
- Third, items were selected and ordered within the test based on discrimination and difficulty analyses. The discrimination index used was the correlation between each individual item and the total score on the TERA-3; items were deemed acceptable for inclusion in the TERA-3 if they demonstrated correlations of .20 or higher. The difficulty index was the percentage of children at any given age who passed each item; items with percentages ranging from 15 to 85 were included in the TERA-3. Of the items that were retained for the final version of the TERA-3, the authors reported median discrimination correlations across ages 3 through 8 ranging from .42 for Form A Conventions items to .53 for Meaning items on both Forms A and B. Median difficulties

across ages ranged from 58 percent of children passing Form A Conventions items to 63 percent children passing Form B Meaning items (see Reid *et al.*, 2001, p. 64).

Criterion-Prediction Validity

Four sets of analyses were conducted for three samples of children to examine the criterion-related validity (termed criterion-prediction validity by the authors, after Anastasi & Urbina, 1997) of the TERA-3. The samples were 1) the standardization sample, 2) a sample of 70 public and parochial school students in the second and third grades in South Dakota, and 3) a Texas sample of 64 second and third grade students with reading disabilities or learning disabilities. Children in each of these samples were tested concurrently with the TERA-3 (both Form A and Form B items) and one or more established measures of reading-related abilities.

The authors indicated that in order for the TERA-3 to be considered a valid test, it should "…correlate well" with these other measures (Reid *et al.*, 2001, p. 67).

- The full standardization sample was assessed concurrently with both the TERA-3 and the previous version of the assessment, the TERA-2 (Reid, *et al.*, 1989). Information on school performance was also collected for a total of 411 children (grades unspecified) in this sample.
    - Correlations between Form A TERA-2 raw composite scores and Form A TERA-3 subtest and Reading Composite raw scores were all high, ranging from .85 for TERA-2 scores with TERA-3 Meaning subtest scores to .98 for TERA-2 scores correlated with TERA-3 Reading Composite scores (see Reid *et al.*, 2001, p. 68).[12]
    - TERA-3 scores were also correlated with teacher judgments about reading-related areas of student achievement, including general reading ability, oral reading, reading comprehension, decoding, spelling, punctuation and capitalization, as well as teacher-reported grades in reading or language arts. These correlations were all moderate to high. Correlations between TERA-3 Form A subtest scores and teacher judgments ranged from .43 between Conventions scores and reading comprehension to .70 between Meaning scores and decoding. Correlations between Reading Quotient scores and teacher judgments ranged from .62 with reading or language arts grades to .70 with both reading comprehension and decoding (see Reid *et al.*, 2001, p. 69).
- Children in the South Dakota sample were assessed with the Stanford Achievement Test Series, Ninth Edition (SAT-9; 1996) as well as the TERA-3 (see Reid *et al.*, 2001, p. 68).
    - Correlations between SAT-9 Total Reading scores and TERA-3 subtest scores were .62, .66, and .41 for Alphabet, Conventions, and Meaning subtests, respectively. SAT-9 Total Reading correlated .57 with TERA-3 Reading Composite scores.
    - SAT-9 Word Study scores correlated .56, .54, and .36 with TERA-3 Alphabet, Conventions, and Meaning subtest scores, respectively. The correlation between Word Study and TERA-3 Reading Composite scores was .38.

[12] The authors presented all validity results for Forms A and B separately. Although there were some differences in the exact correlations, these differences were generally small and not consistently higher with one form versus the other. For purposes of clarity, TERA-3 validity information presented in this section refers to correlations with Form A subtest and Reading Quotient scores.

  - Correlations between SAT-9 Vocabulary scores and TERA-3 Alphabet, Conventions, and Meaning subtest scores were .40, .34, and .42, respectively. SAT-9 Vocabulary scores correlated .47 with TERA-3 Reading Composite scores.
  - SAT-9 Reading Comprehension scores were correlated .66, .66, and .72 with TERA-3 Alphabet, Conventions, and Meaning subtest scores, respectively. The correlation between Reading Comprehension and TERA-3 Reading Quotient scores was .74.
- Children who took part in the Texas study were assessed with the Woodcock Reading Mastery Test-Revised-Normative Update (WRMT-R-NU; Woodcock, 1998) as well as the TERA-3 (see Reid *et al.*, 2001, p. 68).
  - Correlations between WRMT-R-NU Word Identification scores and TERA-3 Alphabet, Conventions, and Meaning subtest scores were .60, .47, and .48, respectively. The correlation between WRMT-R-NU Word Identification and TERA-3 Reading Composite scores was .53.
  - Correlations between WRMT-R-NU Word Attack scores and TERA-3 Alphabet, Conventions, and Meaning subtest scores were .44, .51, and .41, respectively. WRMT-R-NU Word Attack and the TERA-3 Reading Composite had a correlation of .49.
  - Correlations between WRMT-R-NU Word Comprehension scores and TERA-3 Alphabet, Conventions, and Meaning subtest scores were .43, .44, and .51, respectively. WRMT-R-NU Word Comprehension scores were correlated .48 with TERA-3 Reading Composite scores.
  - WRMT-R-NU Paragraph Comprehension scores were correlated .44, .55, and .56 with TERA-3 Alphabet, Conventions, and Meaning subtest scores, respectively. WRMT-R-NU Paragraph Comprehension scores were correlated .60 with TERA-3 Reading Composite scores.
  - WRMT-R-NU Reading Quotient scores were correlated .62, .57, and .61 with TERA-3 Alphabet, Conventions, and Meaning subtest scores, respectively. The correlation between Reading Quotient scores for the two tests was .64.

Construct-Identification Validity

The construct validity of the TERA-3 was assessed by addressing seven testable hypotheses related to the underlying "…constructs presumed to account for test performance…" (Reid *et al.*, 2001, p. 70).

- First, the authors hypothesized that there should be strong correlations between TERA-3 raw scores and child age. In the normative sample, correlations between raw scores and child age were .96 for Alphabet, .95 for Conventions, and .94 for Meaning. No correlation with age was presented for the Reading Composite (Reid *et al.*, 2001, p. 49).
- Second, the authors hypothesized that children who were below average in reading ability could be differentiated from children of average reading ability on the basis of TERA-3 performance. No statistical tests were reported related to this hypothesis; however, mean subtest and Reading Quotient scores for children in the normative sample who were diagnosed as learning disabled, reading disabled, or language impaired were generally lower than the average for the sample as a whole. As indicated earlier, subtests were standardized within age bands to a mean of 10 and a standard deviation of 3, while the

Reading Quotient was standardized to a mean of 100 and a standard deviation of 15. The greatest impairment was evident for the Meaning subtest; children with an identified learning or reading disability had mean scores of 7 on this subtest, while language impaired children had a mean score of 5. For the Conventions subtest, learning or reading disabled children had mean scores of 8, while language impaired children had a mean score of 6. Alphabet subtest means were all within one standard deviation of the normative mean; means were 9, 8, and 10 for learning disabled, reading disabled, and language impaired children, respectively. Mean Reading Quotients were 87 for learning disabled children, 85 for reading disabled children, and 81 for language impaired children (see Reid *et al.*, 2001, p. 72).

- Third, because each of the subtests was designed to tap different but related aspects of reading ability, it was expected that subtests would demonstrate moderate correlations with each other. Correlations for the full normative sample between subtests were .66 for Alphabet correlated with Conventions, .46 for Alphabet correlated with Meaning, and .58 for Conventions correlated with Meaning. These correlations were, as expected, lower than were correlations between Forms A and B of the same subtest (.91, .84, and .90 for Alphabet, Conventions, and Meaning subtests, respectively; see Reid *et al.*, 2001, p. 73).
- Fourth, because reading ability is required for academic success, it was expected that TERA-3 scores would correlate with more general measures of school achievement (i.e., measures that are not specific to reading skills and abilities). This hypothesis was examined with SAT-9 achievement test data from the same South Dakota sample of 70 second and third grade students described above in the section on criterion-prediction validity. SAT-9 Total Mathematics, Problem Solving, Calculation, Language, Environment, Listening, Basic Battery, and Complete Battery scores were correlated with TERA-3 subtest and Reading Composite scores. All correlations were moderate to high. Correlations ranged from .31 for Conventions subtest scores correlated with SAT-9 Calculation scores to .75 for Reading Composite scores correlated with SAT-9 Listening scores (see Reid *et al.*, 2001, p. 74).
- Fifth, TERA-3 performance was expected to be related to measures of general intelligence, particularly those tapping verbal abilities.
    - This hypothesis was again examined with the sample of 70 South Dakota students. In addition to the SAT-9 and TERA-3, these students were assessed with the Otis-Lennon School Ability Test, Seventh Edition (OLSAT-7; Otis & Lennon, 1995). Correlations of TERA-3 subtest and Reading Composite scores with OLSAT-7 Total, Verbal, and Nonverbal scores were moderate to high, ranging from .31 for Meaning subtest scores correlated with OLSAT-7 Verbal scores to .62 for Conventions subtest scores correlated with OLSAT-7 Verbal scores. Associations between TERA-3 scores and OLSAT-7 Verbal scores were not consistently higher than were correlations with OLSAT-7 Nonverbal scores (see Reid *et al.*, p. 75).
    - This hypothesis was further investigated with the Texas sample of 64 children with identified learning or reading disabilities. In addition to the TERA-3 and the WRMT-R-NU (described earlier), these children were assessed with the Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1991). Correlations of TERA-3 Alphabet subtest, Conventions subtest, Meaning subtest, and Reading Composite scores with WISC-III Verbal IQ scores were .79, .74, .77,

and .84, respectively. Correlations with WISC-III Full Scale IQ scores were similar, ranging from .69 for Conventions to .77 for the Reading Composite. Correlations with WISC-III Performance IQ scores were substantially lower, ranging from .37 for Alphabet subtest scores to .42 for Reading Composite scores (see Reid *et al.*, 2001, p. 75).

- Sixth, a confirmatory factor analyses was run to establish whether TERA-3 performance data conformed to the hypothesized three-construct model (i.e., that a model in which Alphabet, Conventions, and Meaning are three separate constructs actually fit the data obtained from the normative sample). Results of the analysis indicated a chi-square with six degree of freedom of 3.08 (p = .005), and a comparative fit index of .999. The authors concluded that their model "provided an outstanding fit to the data…" (Reid *et al.*, 2001, p. 76), and that none of the alternative models tested provided a better fit.
- Seventh, the authors proposed that "Because the items of a particular subtest measure similar traits, the items of each subtest should be highly correlated with the total score of that subtest" (Reid *et al.*, 2001, p. 70). The authors do not provide statistical evidence directly related to this hypothesis; rather, they point to the information on item discrimination discussed above in the summary of content-description validity as supportive evidence for the construct validity of the TERA-3 as well.

.

Reliability/Validity Information from Other Studies

Very few studies have been published about the psychometrics of TERA-3 since its relatively recent publication in 2001.

Comments

- Overall, the information provided on the reliability of the TERA-3 indicates that the test demonstrates good internal consistency, alternative forms and test-retest reliabilities. The test appears to have good reliability at all ages for which it was designed, and is similarly reliable for different ethnic/racial subgroups, for males and females, and for children with reading and learning disabilities and language impairments. Further, the two alternate forms of the measure have similar reliabilities. It should be noted that all children included in both reliability and validity analyses received both Forms A and B of the TERA-3 at the same time – essentially doubling the number of items over what would generally be given in a single assessment. The extent to which this might affect the reported reliabilities or validities of either of the two separate forms is not known.
- Overall, the information provided on the validity of TERA-3 subtest and Reading Composite scores suggest that the measure is a valid assessment of three separate but interrelated reading abilities—knowledge and use of letters (Alphabet), understanding of print conventions (Conventions), and ability to comprehend print (Meaning).
- The TERA-3 appears to be appropriate for a range of subgroups, with studies indicating that the TERA-3 is reliable and valid for children with reading and learning disabilities as well as for children without such disabilities.

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

No examples of studies examining the TERA-3 in relation to environmental variation were found. However a study using the earlier version (TERA-2) considered emergent reading in relation to environmental variation. Hammer, Miccio and Wagstaff (2003) examined differences on TERA-2 scores in a sample of Puerto-Rican children enrolled in Head Start. The authors compared two groups of children—those who had been exposed to both Spanish and English from birth, and those who had been exposed to Spanish from birth but whose first substantial exposure to English was in Head Start. Hammer *et al.* found no significant differences between the two groups in their emergent reading scores as assessed with the TERA-2.

V. Adaptations of Measure

None found.

Woodcock-Johnson III (WJ III) Measures Relevant to Phonological Skills

I. Background Information

Author/Source

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

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 the 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 cognitive abilities tests that assess phonological skills.

As described by the authors

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 90+ 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 (see McGrew & Woodcock, 2001 p. 17).

- Participants were selected using a stratified random sampling design to create a representative sample 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 children (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.

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). For this summary, we focus on four tests from the COG battery that assess phonological skills.

The WJ III COG consists of 20 tests, four of which assess phonological skills (i.e., ability analyze and synthesize speech sounds). The tests measure phonological sensitivity (Test 4: Sound Blending and Test 8: Incomplete Words, administered to individuals ages 2 and older), phonological memory (Test 17: Memory for Words, administered to individuals ages 2 and older), and phonological access (Test 18: Rapid Picture Naming, administered to individuals ages 2 and older). Tests 4 and 8 are part of the standard battery, while Tests 17 and 18 are included in the extended battery (see Mather & Woodcock, 2001b p. 11). The WJ III provides a Phonemic Awareness cluster scale based on scores from Tests 4 and 8. While Tests 17 and 18 tap phonological skills, they are included within the Short-Term Memory, and Cognitive Fluency cluster scales, respectively.

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

Comments
A Phonemic Awareness cluster score can be computed by combining Tests 4 and 8, when scoring by hand. However, the WJ III ACH Test 21, Sound Awareness, must also be administered if one chooses to compute this score by using the available computer programs that assist in the scoring and reporting of WJ III results, the Compuscore and Profiles Program.

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 test. What the respondent is asked to do differs by test. Test 4 (Sound Blending) requires the respondent to listen to a

series of phonemes and blend sounds into a word. Test 8 (Incomplete Words) requires the respondent to listen and identify a word missing a phoneme. Test 17 (Memory for Words) involves presenting the respondent with a series of unrelated words and recording his or her ability to repeat the series. Test 18 (Rapid Picture Naming) requires the respondent to name as many illustrated stimuli as possible within a two-minute period. When using the Compuscore and Profiles Program, Test 21(Sound Awareness) must also be administered in order for it to compute a Phonemic Awareness cluster score. Test 21 requires the child to listen and respond to an auditory stimulus and reports on the child's decoding and spelling ability.

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.)

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

Time Needed and Cost

Time

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

Cost

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

III. Functioning of Measure

Reliability Information from Manual

Internal Reliability

The internal reliabilities forTest 4 (Sound Blending), Test 8 (Incomplete Words) and Test 17 (Memory for Words) were calculated using the split-half procedure. The internal reliability of Test 18 (Rapid Picture Naming) was calculated using the Rasch analysis procedure. Test 4 internal reliabilities were .92 at age 2, .93 at age 3, and .90 at ages 4 and 5. Test 8 internal reliabilities were .92 at age 2, .89 at age 3, .86 at age 4, and .83 at age 5. Test 17 internal reliabilities were .94 at age 2, .90 at age 3, .88 at age 4, and .82 at age 5. Test 18 internal

reliabilities were .91 at age 2, and .98 at ages 3, 4, and 5 (See McGrew & Woodcock, 2001, Appendix A pp. 109-129).

Test-Retest Reliability
Test-retest reliabilities were reported for Test 17(Memory for Words) for 1,196 children and adults (total number for ages 2 to 95). For children ages 2 to 7, the correlation between administrations one to two years apart was .57, and the correlation between administrations three to ten years apart was .77 (see McGrew & Woodcock, 2001 p. 41).

In a separate test-retest study of 165 children and adults in three age groups (7 through 11, 14 through 17, and 26 through 79), Test 19 (Rapid Picture Naming) was administered twice at a one-day time interval in order to control for changes in subjects' traits that might occur across a longer time span. Test-retest reliabilities were adequate for the three age groups: $r = .78$ for ages 7 through 11 and 14 through 17; and $r = .86$ for ages 26 through 79 (see McGrew & Woodcock, 2001 p. 39). The data on test-retest reliability for Test 18 did not include data from children below the age of 6, therefore test-retest reliability is not available for younger children.

No test-retest reliability data were reported for Test 4 (Sound Blending) and Test 8 (Incomplete Words).

Validity Information from Manual
Construct Validity
Construct validity was examined by investigating the patterns of associations among tests and among cluster scores using confirmatory factor analysis. According to McGrew and Woodcock (2001, p. 59-68 and Appendix D, E and F), the expected patterns emerged; tests designed to measure similar constructs were more highly associated than were those measuring widely differing constructs.

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 tests from the WJ III COG and the WJ III ACH that were appropriate for preschoolers. The Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R; Wechsler, 1989) and the Differential Abilities Scale (DAS; Elliott, 1990) were also administered (see McGrew & Woodcock, 2001, p. 69). The authors report concurrent validity analyses involving correlations between WJ III cluster scores and scores on the WPPSI-R and the DAS.

- Correlations between WJ III Phonemic Awareness cluster scores and WPPSI-R IQ scores were .49 (Full Scale IQ), .48 (Verbal IQ), and .43 (Performance IQ).
- Correlations between WJ III Phonemic Awareness cluster scores and DAS scores were .46 (General Conceptual Ability), .45 (Verbal Ability), and .38 (Nonverbal Ability).
- Correlations between WJ III Short-Term Memory cluster scores and WPPSI-R Full Scale IQ scores were .52 (Full Scale IQ), .51 (Verbal IQ), and .43 (Performance IQ). The Short-Term Memory cluster is noted presently because it includes Test 17 (Memory for Words). This test measures both phonological skills and short-term memory. Correlations between WJ III Short-Term Memory cluster scores and DAS scores were .43 (General Conceptual Ability), .41 (Verbal Ability), and .38 (Nonverbal Ability).

- Correlations between WJ III Cognitive Fluency cluster scores and the WPPSI-R Full Scale IQ scores were .48 (Full Scale IQ), .40 (Verbal IQ), and .49 (Performance IQ). The Cognitive Fluency cluster is noted presently because it includes Test 18 (Rapid Picture Naming). This test measures both phonological skills and cognitive fluency.
- Correlations between WJ III Cognitive Fluency cluster scores and DAS scores were .57 (General Conceptual Ability), .45 (Verbal Ability), and .54 (Nonverbal Ability).

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

- Correlations between WJ III Phonemic Awareness cluster scores and SB-IV composite and cluster scores (i.e., Test Composite, Verbal Reasoning, Abstract/Visual Thinking, Quantitative Reasoning, and Short-Term Memory) were examined. Correlations between WJ III Phonemic Awareness and SB-IV clusters ranged from .00 with Abstract/Visual Reasoning to .55 with Short-Term Memory. WJ III Phonemic Awareness cluster scores demonstrated almost identical correlations with Verbal Reasoning (.49), the SB-IV cluster with the greatest apparent similarity to the WJ III Phonemic Awareness cluster, and with SB-IV Quantitative Reasoning (.48).
- Correlations between WJ III Short-Term Memory cluster scores and SB-IV cluster scores ranged from .32 with Quantitative Reasoning to .64 with the SB-IV Test Composite. As might be expected, the SB-IV Short-Term Memory cluster was also one of the most strongly related to the WJ III Short-Term Memory cluster. It had a correlation of .62.
- Correlations between WJ III Cognitive Fluency cluster scores and SB-IV cluster scores ranged from .12 for Quantitative Reasoning to .42 for the SB-IV Test Composite.

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 the previous revision, the WJ-R, but we were unable to find any that are relevant to the preschool age range.

Comments

- Reliability and validity information available in the manual is not always provided at the individual test level. Thus, the psychometric information provided for the phonological skills tests is somewhat limited.
- No information is provided on interrater reliability for the tests of interest here.
- Based on the psychometric information that was available for the WJ III areas relevant to phonological skills, internal reliability appears to be strong.
- Reported concurrent correlations between the WJ III Phonemic Awareness, Short-Term Memory, and Cognitive Fluency cluster scores and three measures of intelligence (i.e., DAS, WPPSI-R, and SB-IV) were moderate to strong. For the Phonemic Awareness cluster, specifically, there was little variability in correlation size across criterion measure scales (e.g., verbal reasoning, quantitative reasoning, general conceptual ability). The one exception to this trend was noted between the WJ III Phonemic Awareness cluster score and the SB-IV Abstract/Visual Reasoning scale, where a null association was found.

That is, the skills tapped by the WJ III phonemic awareness tests seem to require many of the same skills measured by constructs somewhat different than phonemic awareness. This somewhat questions the uniqueness of the underlying construct that these tests measure.

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 tests of the Tests of Achievement (see the WJ III summary included with Math measures section of this review compendium).

V. Adaptations of Measure

Spanish Version of WJ III

A Spanish version of the WJ III is available.

Woodcock-Johnson III (WJ III) Measures Relevant to Print Skills

I. Background Information

Author/Source

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

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 tests related to achievement that assess print skills.

As described by the authors

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 90+ 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 (McGrew & Woodcock, 2001, p. 17).

- Participants were selected using a stratified random sampling design to create a representative sample 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 children (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.

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). For this summary, we focus on three subtests from the ACH battery that assess print skills.

The WJ III ACH consists of 22 tests, at least three of which assess print skills in children under age 6. The tests measure word identification skills (Test 1: Letter-Word Identification, administered to individuals ages 2 and older), writing responses (Test 11: Writing Samples, administered to individuals ages 5 and older), and phonic and structural analysis skills (Test 13: Word Attack, administered to individuals ages 2 and older). Tests 1 and 11 are part of the standard battery, while Test 13 is included in the extended battery (See Mather & Woodcock, 2001a, p. 11).

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

Comments
When achievement tests are used with preschool children, careful consideration should be given to whether the child has had the opportunity to acquire the skills and knowledge that the achievement tests measure.

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 test. What the respondent is asked to do differs by test. Test 1 (Letter-Word Identification) requires the respondent to identify letters and pronounce them correctly. Test 11 (Writing Samples) assesses the respondent's ability to produce written sentences. Test 13 (Word Attack) requires the respondent

to produce sounds for single letters and letter combinations (non-words and low frequency words) presented visually by the assessor.

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 assessment. 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.)

The WJ III is designed to be administered in a one-on-one setting.

Time Needed and Cost

Time

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

Cost

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

III. Functioning of Measure

Reliability Information from Manual

Internal Reliability

The internal reliabilities of Test 1 (Letter-Word Identification) and Test 13 (Word Attack) were calculated using the split-half procedure. The internal reliability of Test 11 (Writing Sample) was calculated using the Rasch analysis procedure. The ages for which internal reliabilities were assessed varied by test. Test 1 reliabilities were .98 at age 2, .97 at age 3, .98 at age 4, and .99 at age 5. Test 11 internal reliability was .70 at age 5, the single preschool age at which it was assessed. Test 13 showed internal reliabilities of .93 at age 4 and .94 at age 5 (See McGrew & Woodcock, 2001, Appendix A pp. 109-129).

Test-retest reliability

Test-retest reliabilities were reported for Test 1 (Letter-Word Identification) for 1,196 children and adults (total number for ages 2 to 95). For children ages 2 to 7, the correlation for two administrations at an interval of less than one year was .96; at an interval of between one to two

years the correlation was .91, and the correlation between assessments between three and ten years apart was .87 (see McGrew & Woodcock, 2001 p. 40).

In a separate test-retest study of 457 children and adolescents ranging in age from 4 to 17 years, all WJ III ACH tests were administered two times with a one-year interval. Test-retest correlations for children ages 4 to 7 were .92 for Test 1 (Letter-Word Identification), .82 for Test 11 (Writing Sample), and .79 for Test 13 (Word Attack; see McGrew & Woodcock, 2001 p. 42).

Validity Information from Manual

Construct Validity

Construct (internal structure) validity was examined by investigating the patterns of associations among tests and among cluster scores using confirmatory factor analysis. According to McGrew and Woodcock (2001, pp. 59-68 and Appendix D, E and F), the expected patterns emerged; tests designed to measure similar constructs were more highly associated than were those measuring widely differing constructs.

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. The Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R; Wechsler, 1989) and the Differential Abilities Scale (DAS; Elliott, 1990) were also administered (see McGrew & Woodcock, 2001, p. 91). The authors reported concurrent validity analyses based on correlations between WJ III test scores and scores in the WPPSI-R and the DAS.

- Correlations between the WJ III ACH Test 1 (Letter-Word Identification) and WPPSI-R IQ scores were .48 (Full Scale IQ), .40 (Verbal IQ), and .47 (Performance IQ).
- Correlations between the WJ III ACH Test 1 (Letter-Word Identification) and DAS scores were .49 (General Conceptual Ability), .39 (Verbal Ability), and .47 (Nonverbal Ability).
- Correlations between the WJ III ACH Test 13 (Word Attack) and WPPSI-R IQ scores were .37 (Full Scale IQ), .35 (Verbal IQ), and .32 (Performance IQ).
- Correlations between the WJ III ACH Test 13 (Word Attack) and DAS scores were .41 (General Conceptual Ability), .46 (Verbal Ability), and .38 (Nonverbal Ability).
- No concurrent validity information was reported for Test 11(Writing Sample).

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

- Reliability and validity information is not provided in the manual for all of the print skills tests for children under age 6. For example, as noted above, there is no concurrent validity information for Test 11 (Writing Sample).
- No information interrater reliability information was provided for this set of tests.

- Each of the tests related to print skills showed strong internal consistency reliability that varied very little by age. As expected, test-retest correlations became smaller over greater amounts of time, yet correlations remained strong even at the largest time intervals between testing.
- Concurrent reliability correlations were generally moderate, suggesting that these WJ III tests are tapping skills that are related to, but not overlapping with, skills assessed by the WPPSI-R and the DAS. This may be appropriate, given that these tests are designed to measure achievement, while the WPPSI-R and the DAS are designed as measures of ability. .
- While the described WJ III ACH tests related to print skills may be administered to preschool children, they may not be the most developmentally appropriate tests to administer to children under the age of 4, especially for children who have not had experience in preschool settings or other environments that provide opportunities to acquire the skills being measured. The most appropriate use of these tests with young children may be to examine the effects of an intervention targeting the specific skills that are measured by the tests.

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 tests of the Tests of Achievement (see the WJ III summary included with the Math measures section of this review compendium).

V. Adaptations of Measure

Spanish Version of WJ III

A Spanish version of the WJ III is available.

References for Literacy Measures

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Developing Skills Checklist. (1990). Riverside, CA: CTB/McGraw-Hill.

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

Elliott, J., Lee, S. W., & Tollefson, N. (2001). A reliability and validity study of the Dynamic Indicators of Basic Early Literacy Skills—Modified. *School Psychology Review, 30,* 33-49.

Good, R. H., Gruba, J., & Kaminski, R. A. (2002). Best practices in using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes-driven model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (4th ed., pp. 699-720). Washington, DC: National Association of School Psychologists.

Good, R. H., & Kaminski, R. A., Eds. (2002a). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu.

Good, R. H., & Kaminski, R. A. (2002b). Nonsense Word Fluency. In R. H. Good & R. A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.), pp. 23-29. Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu.

Good, R. H., & Kaminski, R. A., & Dill, S. (2002). DIBELS Oral Reading Fluency and Retell Fluency. In R. H. Good & R. A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.), pp. 30-38. Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu.

Good, R. H., & Kaminski, R. A., & Smith, S. (2002a). Word Use Fluency. In R. H. Good & R. A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.), pp. 39-43. Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu.

Good, R. H., Kaminski, R. A., & Smith, S. (2002b). Phoneme Segmentation Fluency. In R. H. Good & R. A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.), pp. 16-22. Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu.

Good, R. H., Laimon, D., & Kaminski, R. A., & Smith, S. (2002). Initial Sound Fluency. In R. H. Good & R. A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.), pp. 10-15. Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5,* 257-288.

Good, R. H., Simmons, D., Kame'enui, E., Kaminski, R. A., & Wallin, J. (2002). DIBELS instructional recommendations: Intensive, strategic, and benchmark. In R. H. Good & R. A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.), pp. 48-66. Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu.

Gunn, B., Smolkowski, K., Biglan, A., & Black, C. (2002). Supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school: A follow-up. *Journal of Special Education, 36*, 69-79.

Hammer, C. S., Miccio, A. W., & Wagstaff, D. A. (2003). Home literacy experiences and their relationship to bilingual preschoolers' developing English literacy abilities: An initial investigation. *Language, Speech, & Hearing Services in Schools, 34,* 20-30.

Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review, 32,* 541-556.

Kaminski, R. A., & Good, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25,* 215-228.

Kaminski, R. A., & Good, R. H. (2002). Letter naming fluency. In R. H. Good & R. A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.), pp. 6-9. Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu.

Kamps, D. M., Wills, H. P., Greenwood, C. R., Thorne, S., Lazo, J. F., Crockett, J. L., Akers, J. M., & Swaggart, B. L. (2003). Curriculum influences in growth in early reading fluency for students with academic and behavioral risks: A descriptive study. *Journal of Emotional and Behavioral Disorders, 11,* 211-224.

Karlen, B. K., & Gardner, E. (1985). *Stanford Diagnostic Reading Test, Third Edition*. San Antonio, TX: The Psychological Corp.

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

Marston, D. (1989). Curriculum-based measurement: What is it and why do it? In M. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford.

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 Pyschological Corp.

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

National Reading Panel (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups.* Bethesda, MD: National Institute of Child Health and Human Development. Available: http://www.nationalreadingpanel.org.

National Research Council (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Novack, H. S., Bonaventura, E., & Merenda, P. F. (1973). A scale for early detection of children with learning problems. *Exceptional Children, 40*, 98-105.

Nurss, J. R., & McGauvran, M. E. (1986). *Metropolitan Reading Tests.* San Antonio, TX: The Psychological Corp.

Otis, A. S., & Lennon, R. T. (1995). *Otis-Lennon School Ability Test, Seventh Edition*. San Antonio, TX: Harcourt Assessment.

Reading mastery. (1995). DeSoto, TX: SRA/McGraw-Hill.

Reid, D. K., Hresko, W. P., & Hammill, D. D. (1989). *Test of Early Reading Ability, Second Edition*. Austin, TX: PRO-ED.

Reid, D. K., Hresko, W. P., & Hammill, D. D. (2001). *Test of Early Reading Ability-3*. Austin, TX: PRO-ED.

Shinn, M. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. New York: Guilford.

Stanford Achievement Test Series, Ninth Edition. (1996). San Antonio, TX: Harcourt, Brace, & Company.

Success for all. (1999). Baltimore, MD: Success for All Foundation.

Tindal, G., Marston, D., & Deno, S. L. (1983). *The reliability of direct and repeated measurement* (Research Rep. 109). Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

Torgesen, J. K., & Bryant, B. R. (1994). *Test of Phonological Awareness.* Burlingame, CA: Psychological and Educational Publications.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processing.* Austin, TX: PRO-ED.

Wechsler, D. (1991). *The Wechsler Intelligence Scale for Children, Third Edition*. San Antonio, TX: The Psychological Corp.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence, Revised (WPPSI-R): Manual*. San Antonio, TX: The Psychological Corp.

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

Woodcock, R. W. (1998). *Woodcock Reading Mastery Test, Revised – Normative Update.* Itasca, IL: Riverside Publishing.

Math Measures

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 the author

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 Early Childcare Research Network, 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. Adaptations 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.

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.

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 the authors

"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.

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 area 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.

Stanford-Binet Intelligence Scale, Fourth Edition[13], 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.

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 the authors

"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.

[13] A Fifth Edition of the Stanford-Binet Intelligence Scale was released in 2003, following completion of this profile.

- 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 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, *et al.,* 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, *et al.*, 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, *et al.,* 1986b, p. 46).

Validity Information from Manual
Construct Validity
Correlations were calculated between all subtest, area, and composite scores (see Thorndike *et al.,* 1986b, 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; Wechsler, 1967). 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, *et al*. (1986b) 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, 1986b, 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, *et al.* (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, *et al.* (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.

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

Short Form of the TEMA-2

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).

Woodcock-Johnson III Tests of Achievement (WJ III ACH), Math Subtests

I. Background Information

Author/Source

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

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 the authors

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 include writing a single number, solving simple arithmetic problems, solving word problems read aloud to him/her, and counting and identifying 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.[16] 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);

[16] 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.

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).

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, pp. 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, pp. 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 the previous version of the measure, the WJ-R (Woodcock & Johnson, 1989), 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.

References for Math Measures

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

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

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

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

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 Corp.

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

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

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

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.

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*, 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.

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.

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.

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*, 1072-1077.

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*, 451-477.

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

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 Corp.

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

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 Corp.

Ongoing Observational Measures

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.

Publisher: Teaching Strategies, Inc.
Box 42243
Washington, DC 20015
Phone: 800-637-3652
Website: www.teachingstrategies.com

Purpose of Measure

"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 (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 (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)

- Abbott-Shim (2000) focused on the curriculum component of Creative Curriculum, rather than the assessment component. The study is relevant to the assessment, however, because one of the main purposes of the assessment system is to inform curriculum decisions. 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.

The Galileo System for the Electronic Management of Learning (Galileo)

I. Background Information

Author/Source

Author: Assessment Technology, Inc. (2002) *Galileo online technical manual.* Available: http://www.assessmenttech.com/pages/research/galileotechmanual_files/contents.html.

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.
- 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.

(Note: This profile 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)

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 use of this assessment system requires training in use of the computer program. While training and use do not seem overly difficult, this may be a barrier in some settings.
- 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.
- 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.

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.

V. Adaptations of Measure

Galileo Scale Builder
Galileo Scale Builder allows the system to be translated into other languages.

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.

- *Initiative* includes expressing choices, solving problems, engaging in complex play, and cooperating in routines.
- *Social Relations* includes relating to adults, relating to children, making friends, solving social problems, and expressing feelings.
- *Creative Representation* includes making and building, drawing and painting, and pretending.

- *Music and Movement* includes body and coordination, manual coordination, imitating a beat, and movement and directions.
- *Language and Literacy* includes understanding speech, speaking, interest in reading, using books correctly, beginning reading, and beginning writing.
- *Logic and Mathematics* includes arranging in order, using comparison words, sorting, using the words some, not, and all, comparing numbers, counting objects, spatial relations, and sequence and time.

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) within the

larger construct (e.g., Initiative). For example, indictors for the "Expressing Choices" skill in the "Initiative" category include a) child does not yet express choice to others, b) child indicates a desired activity or place of activity by saying a word, pointing, or some other action, c) child indicates desired activity, place of activity, materials, or playmates with a short sentence, d) child indicates with a short sentence how plans will be carried out, and e) child gives detailed description of intended actions.

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, *et al.*, 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, *et al.*, 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, *et al.*, 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).[17].

V. Adaptations of Measure

None found.

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

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 49 percent male; 8 percent of the children were classified as having special needs; 79 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 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.
- 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 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.
- 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.

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, pp. 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)

Although 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., without random assignment to treatment or control groups), 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.

V. Adaptations of Measure

Spanish Language Versions

There are Spanish language versions of some of the WSS materials.

References for Ongoing Observational Measures

Abbott-Shim, M. (2001). *Validity and reliability of The Creative Curriculum for Early Childhood and Developmental Continuum for Ages 3-5* (Technical Report). Atlanta, GA: Quality Assist.

Abbott-Shim, M. (2000*). Sure Start Effectiveness Study: Final report*. Atlanta, GA: Quality Assist, Inc

Assessment Technology. (2002). *Galileo online technical manual.* Available: 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.

Dichtelmiller, M., Jablon, J., Meisels, S., Marsden, D., & Dorfman, A. (1998). *Using work sampling guidelines and checklists: An observational assessment*. Ann Arbor, MI: Rebus.

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.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research, 59,* 297- 313.

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

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*, 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.

Social-Emotional Measures

Bayley Scales of Infant Development—Second Edition (BSID-II), Behavioral Rating Scale (BRS)

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 the author

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…The [Behavior Rating Scale] assesses the child's behavior during the testing situation, which facilitates interpretation of the Mental and Motor Scales" (Bayley, 1993, p. 1).

The BRS was a substantial revision of the Infant Behavior Record (IBR) that was part of the original BSID. Changes include new and revised items, use of a uniform 5-point rating scale for all items, and greater attention to age-appropriateness of items. Some items are asked only for children within one or two of three age groups: 1- to 5-months, 6- to 12-months, and 13- to 42-months.

Population Measure Developed With

BRS 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 42 months.

Key Constructs of Measure

The BSID-II includes a total of three scales: the Mental Scale, the Motor Scale, and the Behavior Rating Scale (BRS). The focus of this profile is the BRS, which is used by examiners to rate qualitative aspects of children's behavior during assessment sessions from which Mental and Motor Scale scores and age-normed Mental Development Index (MDI) and Psychomotor Development Index (PDI) scores are also derived. There are a total of 30 items on the BRS, not all of which are relevant for all age groups. Two of the items are questions asked of the parent regarding the extent to which the assessment captured the child's typical behavior and provided an adequate representation of the child's skills. Five composites can be constructed from the examiner-report items. Raw scores on each composite have an associated percentile rank within each of the three age groups. Scores at the 25th percentile and higher are classified as *Within Normal Limits*, scores between the 11th and 24th percentile are classified as *Questionable*, and score at or below the 10th percentile are classified as *Non-Optimal*.

- *Attention/Arousal.* This 9-item scale is included for children ages 1- to 5-months only. Items involve arousal, attention, interest and positive and negative affect exhibited in the testing situation.
- *Motor Quality.* The 7 items included in this scale (8 items for children age 13 months and older) involve "…muscle tone, fine and gross motor control, bradykinesia, and the quality of movement" (Bayley, 1993, p. 231).
- *Orientation/Engagement.* This scale is included for children ages 6- to 12-months (11 items) and 13- to 42-months (9 items) and taps approach and avoidance tendencies within the assessment situation. Items involve interest, persistence, enthusiasm, and initiative with tasks in the assessment, and positive affect and engagement with the examiner.
- *Emotional Regulation.* Included for children ages 6- to 12-months (8 items) and 13- to 42-months (10 items), this scale taps characteristics of the child related to the ability to deal with heightened levels of emotion. Items involve negative affect, frustration, hypersensitivity, and hyperactivity as well as attention, cooperation, and adaptation to change.
- *Total Score.* The Total Score is composed of all items for the child's age group, with the exception of two items omitted from the Total Score for the 13- to 42-month age group.

Norming of Measure (Criterion or Norm Referenced)

Norm referenced.

Comments

- This was a major revision of the earlier Infant Behavior Record (IBR). Prior to this revision there was no standardized way of using the IBR, which may have led to an underutilization of information derived from the IBR (Wolf & Lozoff, 1985, cited in Bayley, 1993, p. 21). At the same time, however, several researchers developed IBR factor scores and found that these scores were predictive of children's later functioning (Matheny & Wilson, 1981, cited in Bayley, 1993, p. 21). The BRS was thus designed both to make the measure more useful for examiners and to build upon information developed through independent research demonstrating the predictive validity of examiner's ratings of children's behavior within the testing situation.
- The BRS is designed specifically to be completed following administration of the Mental and Motor components of the BSID-II. Thus, this measure may not be the most efficient choice for assessing socioemotional functioning unless the individualized assessments of cognitive and motor functioning provided by the BSID-II are also desired.

II. Administration of Measure

Who is the Respondent to the Measure?
BSID-II examiner.

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

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
Scoring of the BRS is relatively straightforward. However, 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 all BSID-II test results.

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
According to the Manual, the BSID-II takes approximately 25 to 35 minutes to administer for children under 15-months-old and up to 60 minutes for children older than 15-months. No specific time for completion of the BRS was indicated by Bayley (1993). The measure is

relatively short (30 items, not all of which are relevant for any single age group) and should take no more than 10 minutes to complete.

Cost

- Complete BSID-II kit: $950
- Manual: $80

III. Functioning of Measure

Reliability Information from Manual

Internal Reliability

Internal reliability estimates (coefficient alphas) were computed within multiple 1-month age groups (see Bayley, 1993, p. 191).

- Alphas for the Attention/Arousal composite, which is only scored for children between 1 and 5 months of age, ranged from .64 at 2 months of age to .82 at 1 month of age, with an average alpha of .74.
- Alphas for Orientation/Engagement, scored for children between 6 and 42 months of age, ranged from .83 at 27 months to .90 at 42 months, with an average alpha of .87.
- Emotional Regulation is scored for children between 6 and 42 months of age. Alphas for this composite ranged from .73 at 8 months to .90 at 24 months. The average alpha was .84.
- Alphas for Motor Quality, scored for all children, ranged from .74 at 24 months to .86 at both 2- and 21-months. The average alpha was .82.
- Total Score alphas ranged from .82 at 2-months to .92 at 42-months. The average alpha was .88.

Test-Retest Stability

The BSID-II was administered twice to a sample of 175 children, drawn from four age groups in the standardization sample (1, 12, 24, and 36 months). Children were re-tested between 1 and 16 days after their first assessment, with a median interval of 4 days.

- At 1 month of age, test-retest correlations were .55 for the BRS Total Score, .48 for Attention/Arousal, and .70 for Motor Quality (see Bayley, 1993, p. 193).
- At 12 months of age, test-retest correlations were .90 for the Total Score, .57 for Orientation/Engagement, .69 for Emotional Regulation, and .86 for Motor Quality (see Bayley, 1993, p. 193).
- Test-retest correlations for the combined 24- and 36-month age groups were .60 for the Total Score, .61 for Orientation/Engagement, .66 for Emotional Regulation, and .71 for Motor Quality (see Bayley, 1993, p. 194).
- As an additional measure of test-retest reliability, the percentages of children who received the same classification (Within Normal Limits versus the combined Questionable and Non-Optimal classifications) in the two assessments were examined (see Bayley, 1993, p. 195).
    - At age 1-month, 80.9 percent of children were similarly classified for Attention/Arousal, 90.5 percent had the same classification based on Motor Quality, and 73.3 percent received the same classification based on their Total Scores.

- At age 12-months, the percentages of children who received the same classification based on Orientation/Engagement, Emotional Regulation, Motor Quality, and Total Scores were 85.3, 83.3, 93.7, and 87.5 percent, respectively.
- At ages 24-and 32-months, the percentages of children who received the same classification based on Orientation/Engagement, Emotional Regulation, Motor Quality, and Total Scores were 94.1, 89.4, 96.5, and 90.6 percent, respectively.

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 each assessment from nearby and who also rated the children on the Mental and Motor Scales). Bayley (1993, p. 196) reports interrater correlations for the youngest (2- to 5-months) and oldest (13-to 30-months) age groups. In the youngest age group, interrater correlations were .70 for the BRS Total Score, .57 for Attention/Arousal, and .80 for Motor Quality. In the oldest age group, interrater correlations were .88 for the Total Score, .82 for Orientation/Engagement, .83 for Emotional Regulation, and .79 for Motor Quality. Bayley also examined whether children would receive the same classification (Within Normal Limits versus combined Questionable and Non-Optimal classifications) based on the two raters' evaluations (see Bayley, 1993, p. 196). Classification agreement was 90.9 percent for all factors (Total Score, Attention/Arousal, and Motor Quality) in the 1 to 5 month age group. In the 13- to 42-month age group, agreement in classification was 87.5 percent for the Total Score and for Emotional Regulation, 90.0 percent for Orientation/Engagement, and 95.0 percent for Motor Quality.

Validity Information from Manual

Construct Validity

Bayley (1993) reported a series of exploratory factor analyses with both the standardization sample and a clinical sample of children who had participated in special group studies. Included in this group were 57 premature children, 35 HIV positive children, 137 drug exposed children, 25 children who had asphyxia at birth, 60 children with Down Syndrome, 14 children with identified developmental delays related to medical complications, and 22 children with chronic otitis media (ear infections). Analyses were conducted separately for children ages of 1- to 5-months, 6- to 12-months, and 13- to 42-months. According to Bayley (p. 207), "The number of factors to be extracted was identified through the use of an integrated rational approach involving factor structure interpretability, subjective examination of the scree plot, and…(eigenvalue greater than or equal to 1.0)." Results of these analyses indicated two factors (Motor Quality and Attention/Arousal) for the youngest age group, and three factors (Orientation/Engagement, Motor Quality, and Emotional Regulation) for the two older age groups. The pattern of factor loadings on each of the factors was fairly (although not entirely) consistent across the clinical and standardization samples.

Criterion-Related Validity

Bayley (1993) reported correlations between BRS scores and scores on the BSID-II Mental and Psychomotor Development Indices (MDI and PDI) within the standardization sample (see p. 213).

- In the 1- to 5-month age group, correlations with MDI scores were .40 for Attention/Arousal, .28 for Motor Quality, and .38 for Total scores. Correlations with PDI scores were .25 for Attention/Arousal and .27 for both Motor Quality and Total scores.
- Correlations with MDI scores in the 6- to 12-month age group were .46 for Orientation/Engagement, .26 for Emotional Regulation, .33 for Motor Quality, and .45 for Total scores. Correlations with PDI scores were .30 for Orientation/Engagement, .13 for Emotional Regulation, .37 for Motor Quality, and .33 for Total scores.
- In the 13- to 42-month age group, correlations with MDI scores were .34 for Orientation/Engagement, .33 for Emotional Regulation, .20 for Motor Quality, and .37 for Total scores. Correlations with PDI scores were .23 for Orientation/Engagement, .22 for Emotional Regulation, .18 for Motor Quality, and .27 for Total scores.

Bayley (1993, p. 214) also examined the extent to which children with BRS scores within the Non-Optimal range also fell within the MDI and PDI Significantly Delayed ranges (i.e., scores below 70). These analyses were conducted with the standardization sample combined with the clinical sample.

- Looking at BRS Total Scores, 78.5 percent of children ages 1- to 5-months in the total sample were correctly classified on the MDI (i.e., had both a Non-Optimal BRS score and an MDI score within the Significantly Delayed ranged, or had both a Within Normal Limits or Questionable BRS score and an MDI score of 70 or higher). In the 6- to 12-month age group this percentage was 66.3 percent; in the 13- to 42-month age group, 69.8 percent were correctly classified. Percentages of correct MDI classification based on the four BRS scale scores (Motor Quality, Arousal/Attention, Orientation/Engagement, and Emotional Regulation) ranged from 60.7 to 72.6 percent across the three age groups.
- Percentages of correct classification on the PDI, based on the BRS Total Scores were 67.7 percent in the 1- to 5-month age group, 72.5 percent in the 6- to 12-month age group, and 66.5 percent in the 13- to 42-month age group. Percentages of correct PDI classification based on the four BRS scales ranged from 57.2 to 77.2 percent across the three age groups.

Reliability/Validity Information from Other Studies

Thompson and colleagues (Thompson, Wasserman, & Matula, 1996) conducted a series of principal components analyses similar to those reported by Bayley (1993), using data from subsamples of the BSID-II standardization and clinical samples. In total, there were 1,341 children from the standardization sample and 765 children from the clinical samples included in these analyses. Within the three age groups (1 to 5 months, 6 to 12 months, and 13 to 42 months), analyses were conducted for each of the two subsamples separately, and for the combined standardization and clinical subsamples. Results of first- and second-order factor analyses generally paralleled those reported by Bayley, suggesting two factors for the 1- to 5-month age group, and three factors for the older age groups. In the first order factor analyses that most directly paralleled those reported in the manual, the factors were generally consistent with those identified by Bayley (Motor Quality and Attention for the 1- to 5-month age group, and Motor Quality, Orientation/Engagement, and Emotional Regulation for the older age groups). Although the factors were fairly consistent in the second order factor analyses as well, there were some exceptions. First, in the 1- to 5-month age group, a clear Motor Quality factor emerged in

the clinical and combined samples analyses, but not the standardization sample analyses. In the 6- to 12-months age group, a clear Emotional Regulation factor was evident only for the combined samples analysis. In the 13- to 42-month age group, a clear Emotional Regulation factor was evident for the standardization and combined samples, but not for the clinical sample.

Comments

- Information provided by Bayley (1993) on the reliability of the BRS indicates that the scales demonstrated high internal consistency in the standardization sample with only one exception; internal consistency of the Attention/Arousal composite was moderate for 2-month-olds. Further, test-retest correlations of scale scores and Total scores were moderate to high across a short time span, and correlations were moderate to high between ratings made by an examiner and by a second rater who observed the testing sessions. Taken together, these findings provide support for the reliability of the BRS.
- Test-retest reliability information based on classifications, rather than scores, appears to indicate that consistency of being in the Within Normal Limits range, versus the combined Questionable and Non-Optimal ranges, tends to increase across the three age groups. Indeed, more than one-fourth of all 1-month-old children assessed actually changed classification based on their Total Scores. There were several limitations to the information presented in the manual, however, that may limit conclusions that can be made.
    - These were fairly small samples, and the 1-month and 12-month groups were approximately half the size (42 and 48 children, respectively) of the combined 24 and 36-month groups (85 children).
    - Bayley (1993) does not report whether there were any consistent patterns in the direction of change, or whether positive and negative changes were equally likely. It might be particularly interesting to know whether movement was greater among those in the Questionable range than among other children.
- As with information on test-retest reliability of classifications, there were several limitations to information presented in the manual on interrater reliability based on classifications.
    - The sample was small, and there was no information on the number of children in each of the age groups.
    - Information was provided only for the youngest and oldest age groups, and not for children in the middle age group (between 6 and 12 months). The reason for this was not explained.
- With regard to validity of the BRS, Bayley (1993, p. 214) concludes, "The data presented [in the manual] suggest that the BRS demonstrates evidence of content, construct, and criterion-based validity" but also adds that "The evidence presented here should be viewed as an initial effort to explore the validity of the BRS." There are several limitations to the validity analyses presented in the manual that may require further research.
    - The majority of validity analyses reported in the manual address the extent to which BRS scores and classifications are associated with MDI and PDI scores and classifications obtained from the same testing session. Additional information regarding prediction across time, as well as prediction from the BRS to other measures of similar constructs would be useful to further address questions of the

validity of the BRS for any use other than to facilitate clinical interpretation of children's MDI and PDI scores.

- Most of the analyses that were presented, particularly the factor analyses and correlations between the BRS scale scores and MDI and PDI scores, were presented without any discussion of *a priori* hypotheses. Thus, whether the findings provide strong or modest support for the validity of the measure is not clear.
- It is worth noting that the Questionable range was combined with the Non-Optimal range in the reliability analyses, while in the reported validity analyses, it was combined with the Within Normal Limits range. There is no explanation provided for either choice.

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

None found.

Comments

Several studies were found examining associations between environmental variations and children's scores on a previous version of the BRS – the IBR (e.g., Hans & Jeremy, 2001). However, the BRS was a major revision of the IBR, including changes in items and changes in item response scales. Further, the IBR did not have a standard scale structure and different researchers constructed scales differently. Therefore, the relevance of these studies to the more standardized BRS is unknown.

V. Adaptations of Measure

None found.

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 the authors

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).

  - *Social 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, 1989a), 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, pp. 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, 1989b), 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 *et al.* 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 *et al.* (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 *et al.* (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.

Child Behavior Checklist/1½ -5 (CBCL/1½-5) and Caregiver-Teacher Report Form (C-TRF)

I. Background Information

Author/Source

Authors: 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.

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 the authors

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, pp. 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, pp. 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 Wakschlag (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 General Cognitive Index of the McCarthy Scales of Children's Abilities (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, Skuban, & Horwitz, 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 (National Center for Health Statistics, 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).

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 the author

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, pp.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
        - Alphas for the factor analytically derived scales ranged from .76 to .90 (median = .84) for males and .75 to .88 (median = .86) for females.
        - Alphas for Auxiliary scales ranged from .69 to .94 (median = .89) for males and .77 to .91 (median = .86) for females.
    - Six to Eight Year Olds
        - Alphas for factor analytically derived scales ranged from .75 to .93 (median = .85) for males and .81 to .93 (median = .91) for females.
        - Alphas for Auxiliary scales 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
        - Alphas for the factor analytically derived scales ranged from .84 to .95 (median = .87) for males and .59 to .83 (median = .80) for females.
        - Alphas for the Auxiliary scales ranged from .78 to .96 (median = .93) for males and .74 to .87 (median = .82) for females.
    - Six to Eight Year Olds
        - Alphas for the factor analytically derived scales ranged from .82 to .94 (median = .91) for males and .84 to .93 (median = .91) for females.
        - Alphas for the Auxiliary scales 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, pp.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, pp.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, pp. 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, 1995) 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, pp.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, pp.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.

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 Concern*. 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?
N/A.

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.

Infant-Toddler Social and Emotional Assessment (ITSEA)

I. Background Information

Author/Source

Source: Carter, A. S., & Briggs-Gowan, M. J. (2001). *Infant-Toddler Social and Emotional Assessment (ITSEA). Manual, Version 1.1.* Unpublished manual.

Carter, A. S., Briggs-Gowan, M. J., Jones, S. M., & Little, T. D. (2003). The Infant Toddler Social and Emotional Assessment (ITSEA): Factor structure, reliability, and validity. *Journal of Abnormal Child Psychology, 31,* 495-514.

Publisher: Unpublished. Manual, instruments, and supporting articles are available from the authors. Email contact: ITSEA@yale.edu

Purpose of Measure

As described by the authors

The ITSEA was designed to be "…a developmentally and clinically sensitive measure for use with parents and caregivers…The ITSEA was developed to assess social-emotional problems and competencies…." in infants and toddlers (Carter & Briggs-Gowan, 2001, p. 3).

Population Measure Developed With

- Psychometric work and norms development for the ITSEA were conducted with a Community Survey sample of infants and toddlers. All children were born at Yale-New Haven Hospital between July, 1995, and September, 1997. Based on birth records information (e.g., birth weight, gestational age, APGAR scores, presence of chromosomal anomalies such as Down Syndrome, and anoxia at birth), children who were expected to have substantial developmental delays were excluded from the potential participant pool, as were children who were adopted. Only one child per mother was included. An age and sex stratified random sample of 1,788 families was selected from the 7,433 families determined to be eligible for inclusion. Of these 1,788, additional families were excluded if 1) neither parent spoke English well enough to complete the ITSEA either by themselves or as an interview, 2) parents had lost custody of the child, or 3) the family had moved out of state. Of the remaining 1,605 families, 1,280 agreed to participate in the study.
- The final sample included 49 percent boys and 51 percent girls; approximately 47 percent of the children were in the 12- to 23-month age range, 52 percent were in the 24- to 36-month age range, and approximately 4 percent were between 37 and 42 months of age. Approximately 66 percent of the children were white, 17 percent were black, 8 percent were Hispanic, 3 percent were Asian, and 6 percent were multiracial minority.
- Most respondents (96 percent) were biological mothers, with the remainder being biological fathers (approximately 3 percent) or grandmothers or other female guardians (less than 1 percent). This was a fairly well-educated sample, with 41 percent of mothers and 44 percent of fathers having a college degree or more, 32 percent of mothers and 27 percent of fathers having some education beyond high school, 18 percent of mothers and

24 percent of fathers having a high school degree or GED, and 8 percent of mothers and 5 percent of fathers having less than a high school education. The majority of families (66 percent) had incomes putting them at 185% of the poverty line or higher; 16 percent were living in borderline poverty (between the poverty line and 185% of poverty), and 18 percent of the families were living below the poverty line.

Age Range Intended For

Ages 12 months through 48 months.

Key Constructs of Measure

There are several different types of scores that can be derived from the ITSEA. There are four broad domain scores. Within each behavioral domain, items are clustered into more specific behavioral scales for which scores can also be derived.

- *Externalizing Symptoms* (24 items). This domain-level scale includes items from three behavioral scales.
  - *Activity/Impulsivity* (6 items). Activity level, restlessness, tendency to get "wound up" while playing, and tendency to get hurt.
  - *Aggression* (12 items). Disobedience, temper tantrums, aggression towards parents and animals, destructiveness.
  - *Peer Aggression* (6 items). Aggression towards peers, including both overtly physical (e.g., hitting, kicking, biting) and less physical forms (e.g., teasing, excluding, bullying).
- *Internalizing Symptoms* (30 items). There are four behavioral scales included within this domain.
  - *Depression/Withdrawal* (9 items). Sadness, lack of energy, negative self-feelings, and avoidance of contact with others.
  - *General Anxiety* (10 items). Specific and nonspecific fears and anxieties, including several items associated with obsessive compulsive symptoms (e.g., perfectionism, worries about getting dirty, need for order).
  - *Separation Distress* (6 items). Distress at separations as well as needs for attention and physical contact.
  - *Inhibition to Novelty* (5 items). Shyness, slowness to "warm up" to new places or people.
- *Dysregulation* (34 items). There are four behavioral scales included within this domain.
  - *Sleep* (5 items). Difficulties going to sleep alone, and staying asleep.
  - *Negative Emotionality* (13 items). Irritability, anger, crying, difficulty calming down or being soothed, easily upset or frustrated.
  - *Eating* (9 items). Pickiness, food refusal.
  - *Sensory Sensitivity* (7 items). Extreme sensitivity to visual, auditory, vestibular, tactile, and olfactory stimulation.
- *Competence* (37 items). Six behavioral scales are included in this domain.
  - *Compliance* (8 items). Following rules, complying with requests.
  - *Attention* (5 items). Ability to focus attention on a toy, a book, or a teaching situation.
  - *Imitation/Play* (6 items). Imitating sounds or actions, pretend play.
  - *Mastery Motivation* (6 items). Curiosity, enjoyment of challenging activities.

- *Empathy* (7 items). Awareness of others' feelings, attempting to help others in distress.
- *Prosocial Peer Relations* (5 items). Playing well with other children.

In addition, there are three index scores that are not considered scales because items are not necessarily expected to be highly correlated with each other. These indices include items that are of particular clinical significance.

- *Maladaptive Index* (13 items). Symptoms of Tourette Syndrome, Post-Traumatic Stress Disorder (PTSD), toileting problems, sexualized behavior, and Pica (eating or drinking inedible substances).
- *Social Relatedness Index* (10 items). Interest in others, social responsiveness, displaying affectionate behavior toward others. This index and the Atypical Behaviors Index (described below) are included "…to assess behaviors that may be indicative of the presence of [Pervasive Developmental Disorder]/Autism" (Carter & Briggs-Gowan, 2001, p. 3)
- *Atypical Behaviors Index* (8 items). Repetitive movements or activities, lack of awareness of environment, lack of pointing to ask for or indicate something.

Finally, there are 10 items that are not included on any index or scale. These items describe behaviors that are infrequent in the population but that are of substantial clinical significance when they do occur. These include, for example, self-destructive behavior and lack of fear.

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

Comments

- The sample used for the analyses described in the manual, including the development of standardized (*T*) scores and determination of clinically relevant cutpoints was not a nationally representative sample. Indeed, all children were born in the same hospital and resided within the same geographical area. As the authors caution, "…the Greater New Haven community is not representative of the U.S. population, T-scores may not generalize to all communities in the U. S." (Carter & Briggs-Gowan, 2001, p. 10). Additional studies examining the psychometric properties of this instrument with national samples or samples from other regions of the U. S. would be useful.
- The original sample, as indicated above, included children who were predominantly between the ages of 12 and 36 months of age; a small percentage were between 36 and 42 months of age. The ITSEA has since been normed for use with children up to 48 months of age.

II. Administration of Measure

Who is the Respondent to the Measure?
Parents or other caregivers.

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

Who Administers Measure/Training Required?
Test Administration
This measure is typically administered as a questionnaire, and little specific training is required. It can also be administered as an interview (e.g., if the parent or caregiver has limited English reading ability). Interviewers should be well trained to administer the ITSEA in a standard manner, reading all questions verbatim and without offering interpretations of items or other assistance.

Data Interpretation
Carter and Briggs-Gowan (2001, p. 3) indicate that ITSEA should be "…interpreted by individuals who have an understanding of standardized assessment tools and have been trained to discuss the results and limitations of such instruments with parents." Such professionals may include, for example, psychologists, psychiatrists, other mental health professionals, social workers, pediatricians, and nurse practitioners.

Setting (e.g. one-on-one, group, etc.)
This assessment is designed to be administered in a one-on-one or independent (i.e. parent-report) setting.

Time Needed and Cost
Time
According to the Manual, the ITSEA takes approximately 20 to 30 minutes to complete independently as a questionnaire, or 35 to 45 minutes to complete in an interview format.

Cost
There is no cost at this time. All materials are available directly from the authors.

Comments
The ITSEA requires a fourth- to sixth-grade reading level to complete. But in initial work most parents with less than a high school education chose to self-administer the ITSEA as a questionnaire, even when they were offered to have it read to them.

III. Functioning of Measure

Reliability Information from Manual
Internal Reliability
Internal reliability estimates (coefficient alphas) were computed for the full age range (see Carter & Briggs-Gowan, 2001, pp. 27-33).

- The alpha for the Externalizing Symptoms domain scale was .87. Alphas for the three behavioral scales included in this composite were .73, .79, and .79 for Activity/Impulsivity, Aggression, and Peer Aggression, respectively.

- The Internalizing Symptoms domain scale had an alpha of .80. Alphas for the behavioral scales were as follows: Depression/Withdrawal, .74; General Anxiety, .71; Separation Distress, .73; Inhibition to Novelty, .77.
- The alpha for Dysregulation was .86. The four behavioral scales, Sleep, Negative Emotionality, Eating, and Sensory Sensitivity had alphas of .78, .84, .78, and .63, respectively.
- The Competence domain scale had an alpha of .90. Alphas for the six behavioral scales were as follows: Compliance, .74; Attention, .70; Imitation/Play, .59; Mastery Motivation, .62; Empathy, .82; and Prosocial Peer Relations, .66.
- Alphas for the Maladaptive Index, the Social Relatedness Index, and the Atypical Behaviors Index were .56, .56, and .45, respectively. As noted earlier, the items on these three indices were not expected by the authors to be highly correlated with each other.

Test-Retest Reliability

Test-retest reliability was examined with a sample of 93 parents who participated in a Methodological Substudy. This sample was derived from the Community Survey sample and a separate Early Intervention sample of parents who had sought evaluation or services for their children. Carter and Briggs-Gowan (2001) indicated that this sample was similar to the Community Survey sample with respect to parent age, education level, marital status, and ethnicity. Parents in this sample completed the ITSEA twice. Time between assessments ranged from 11 to 44 days, with an average interval of 27 days. Of the 93 families, 41 parents completed the ITSEA as a questionnaire at both administrations, 49 completed the ITSEA first as a questionnaire and then as an interview, and method of administration at the second time point was unknown for the remaining three families. The 90 families for whom information was available regarding method of administration were also included in cross-administration analyses (see Carter & Briggs-Gowan, 2001, p. 56 for test-retest analysis results, p. 57 for cross-administration analysis results).

- The intraclass correlation for the Externalizing Symptoms domain scale was .82. Correlations for the three behavioral scales were .83 for Activity/Impulsivity, .85 for Aggression, and .69 for Peer Aggression. When examining correlations for parents who completed a questionnaire on both occasions (the Questionnaire-Questionnaire group) and those who were interviewed on the second occasion (the Questionnaire-Interview group) separately, all correlations were .80 or higher with one exception: The test-retest correlation for Peer Aggression was .57 for the Questionnaire-Questionnaire group. According to the authors, the difference in correlations between the two groups on this measure (.57 versus .80 for the Questionnaire-Interview group) was not significant (see p. 55).
- The Internalizing Symptoms domain scale had an intraclass correlation of .83. Correlations for the behavioral scales were .74 for Depression/Withdrawal, .85 for General Anxiety, .80 for Separation Distress, and .76 for Inhibition to Novelty. Test-retest correlations for the Questionnaire-Questionnaire and Questionnaire-Interview groups ranged from .70 to .90, with one exception: the intraclass correlation for Depression/Withdrawal was .49 for the Questionnaire-Interview group (versus .80 for the Questionnaire-Questionnaire group). Again, no group differences were reported as significant.

- The intraclass correlation for Dysregulation was .91. Correlations for the four behavioral scales, Sleep, Negative Emotionality, Eating, and Sensory Sensitivity were .88, .85, .84, and .82, respectively. Intraclass correlations for the Questionnaire-Questionnaire and Questionnaire-Interview groups ranged between .76 and .92, and no group differences were significant.
- The Competence domain scale had an intraclass correlation of .90. Alphas for the six behavioral scales were .77 for Compliance, .78 for Attention, .88 for Imitation/Play, .83 for Mastery Motivation, .84 for Empathy, and .80 for Prosocial Peer Relations. The range of correlations for the Questionnaire-Interview and Questionnaire-Questionnaire groups was .70 to .93, and no group differences were significant.
- No intraclass correlations were reported for the index scores (i.e., Maladaptive Index, Social Relatedness Index, and Atypical Behaviors Index).

Interrater Agreement

Interrater agreement was examined through intraclass correlations between mothers' and fathers' reports for a subsample of 100 children who took part in the Methodological Substudy (see Carter & Briggs-Gowan, 2001, p. 56).

- The intraclass correlation for the Externalizing Symptoms domain scale was .69. Intraclass correlations for the three behavioral scales were .65 for Activity/Impulsivity, .69 for Aggression, and .73 for Peer Aggression.
- The intraclass correlation for the Internalizing Symptoms domain scale was .58. Correlations for the behavioral scales were .43 for Depression/Withdrawal, .64 for General Anxiety, .53 for Separation Distress, and .53 for Inhibition to Novelty.
- The intraclass correlation for Dysregulation was .79. Interrater agreements for the four behavioral scales were as follows: Sleep, .78; Negative Emotionality, .73; Eating, .76; and Sensory Sensitivity, .66.
- The intraclass correlation for the Competence domain scale was .76. Intraclass correlations for the six behavioral scales were .56 for Compliance, .71 for Attention, .71 for Imitation/Play, .47 for Mastery Motivation, .71 for Empathy, and .73 for Prosocial Peer Relations.
- No interrater agreement analyses were reported for the three index scores.

Validity Information from Manual

Criterion and Construct Validity

Construct and criterion validity were assessed by examining associations between ITSEA scale scores and additional measures thought to tap the same or related constructs. Three types of measures were examined: 1) parent-report checklists of child problem behaviors and temperament, 2) developmental assessments, and 3) ratings of child problems by early intervention service providers.

- *Parent-report checklists*. Correlations were reported between ITSEA domain scale scores and scores on the Child Behavior Checklist, 2-3 year-old version (CBCL/2-3; Achenbach, 1992), the Parenting Stress Index (PSI; Abidin, 1990), and the Colorado Child Temperament Inventory (CCTI; Buss & Plomin, 1975).
    - In a subsample of 624 two-year-old children from the Community Sample, ITSEA Externalizing domain scores correlated .71 with the CBCL/2-3 Externalizing scale and also correlated .52 with the CBCL/2-3 Internalizing scale

and .67 with the Total scale. ITSEA Internalizing domain scores correlated .57 with CBCL/2-3 Internalizing, .33 with CBCL/2-3 Externalizing, and .48 with CBCL/2-3 Total Problems. Dysregulation domain scores from the ITSEA correlated .52 with CBCL/2-3 Internalizing, .49 with CBCL/2-3 Externalizing, and .61 with CBCL/2-3 Total scores. Finally, ITSEA Competence scores correlated -.28, -.34, and -.31 with CBCL/2-3 Internalizing, Externalizing, and Total scores, respectively (see Carter & Briggs-Gowan, 2001, p. 64). Correlations between CBCL/2-3 scores and ITSEA Index scores were not reported; however the authors indicate that correlations between Maladaptive and Atypical Index items and CBCL/2-3 Internalizing and Externalizing scores ranged from .15 to .42 (see Carter & Briggs-Gowan, p. 62).

- Carter and Briggs-Gowan (2001, p. 62) stated that correlations between ITSEA scales and CCTI measures were expected to be significant but not as high as correlations between ITSEA and CBCL/2-3 scales, due to the lesser overlap in the constructs being assessed. As expected, correlations between ITSEA Externalizing, Internalizing, Dysregulation, and Competence scores and CCTI Sociability, Emotionality, and Soothability scores were all significant and in the expected direction (i.e., scales expected to correlate positively were positively correlated, while scales expected to correlate negatively were negatively correlated), but were generally somewhat lower than correlations of ITSEA scales with CBCL/2-3 scales. Correlations ranged in magnitude from -.10 between ITSEA Externalizing and CCTI Sociability to -.57 between ITSEA Internalizing and CCTI Sociability (see p. 65).
- Correlations between ITSEA Internalizing, Externalizing, and Dysregulation scores and PSI Difficult Child, Parent-Child Dysfunction, Parental Distress, and Total scores ranged from .21 for both Internalizing and Dysregulation correlated with Parent-Child Dysfunction to .48 for Externalizing with Difficult Child (see Carter & Briggs-Gowan, 2001, p. 64) in a sample of 1,225 families from the Community Sample.

- *Developmental assessments*. Associations of ITSEA scores with scores on two developmental assessments were reported—the Vineland Adaptive Behavior Scales (Sparrow, Balla, & Cicchetti, 1984), and the Mullen Scales of Early Learning (Mullen, 1989; see Carter & Briggs-Gowan, 2001, p. 66). A total of five Vineland scores (Communication, Daily Living Skills, Socialization, Motor Skill, and an Adaptive Behavior Composite) and seven scores derived from the Mullen Scales (Expressive Language, Receptive Language, Visual Recognition, Fine Motor, Gross Motor, Cognitive T-Score, and Early Learning Composite) were examined. These assessments were conducted in homes with a sample of 154 children. Carter and Briggs-Gowan expected that the strongest associations would be between the ITSEA Competence domain score and each of the Mullen and Vineland scores, due to the "…strong association between competence and chronological age, which we believe is a marker for mental age" (p. 62).
  - Correlations between ITSEA Competence scores and the three composite scores from the two developmental tests were .41, 45, and .44 with the Vineland Adaptive Behavior Composite, the Mullen Cognitive T-Score, and the Mullen Early Learning Composite, respectively. Correlations between ITSEA Competence scores and all other scale scores from the Vineland and Mullen

assessments ranged from .06 (*ns*) with the Mullen Gross Motor scale to .57 with the Vineland Communication scale. In total, 10 of 12 correlations were significant.
    - There were no significant correlations between ITSEA Internalizing scale scores and scores on either the Mullen or the Vineland scales and composites.
    - The ITSEA Dysregulation scale had a single significant correlation with a Vineland scale (-.16 with Vineland Socialization) and no significant correlations with any Mullen scales or composites.
    - For ITSEA Externalizing, 9 of 12 correlations with Vineland and Mullen scales and composites were significant. Correlations ranged from .01 with Mullen Gross Motor to -.30 for Vineland Communication.
    - 7 of 12 correlations between an ITSEA Social Relatedness/Atypical Index composite and Mullen and Vineland scores and composites were significant. Correlations ranged from -.03 with Mullen Fine Motor to -.33 with the Vineland Adaptive Behavior Composite.
- *Ratings by early intervention providers*. To obtain a measure of criterion validity, 157 children who were part of the Methodological Substudy were rated by evaluators, described by the authors as "early intervention providers" (Carter & Briggs-Gowan, 2001, p. 63), who observed the children during 2-hour home visits. Specific behavior problems similar to those addressed in the parent-report ITSEA were included, with evaluators rating the presence or absence of problems in each area on a 4-point scale ranging from "Not a Problem" to "Definite Problem," while specific competencies were rated on a 7-point scale ranging from "Definite Weakness" to "Definite Strength." Correlations between evaluator ratings and parent ITSEA ratings for the same domain were .39 for Externalizing, .19 for Internalizing, .45 for Dysregulation, .58 for a Social Relatedness/Atypical Index composite, and .40 for Competence (see p. 67).

Reliability/Validity Information from Other Studies

None found. Carter, Briggs-Gowan, and colleagues conducted all reliability and validity analyses currently available with the same samples of subjects. Carter, Briggs-Gowan, Jones, and Little (2003) revised CBCL/2-3 scoring to conform to a newer version of the CBCL designed to span the full preschool period, the CBCL/1.5-5 (Achenbach & Rescorla, 2000). Only very minor changes were found in correlations between CBCL/2-3 and ITSEA measures.

Comments

- Information provided in the manual on the reliability of the ITSEA scales generally indicates good internal consistency, test-retest reliability (including cross-administration reliability), and good interrater reliability.
- The validity of the ITSEA was supported by the reported pattern of correlations with CBCL/2-3 scales. There was a relatively high correlation between ITSEA Externalizing and CBCL/2-3 Internalizing, which may suggest limited discriminant validity. However, the correlation between ITSEA Internalizing and Externalizing scales was a moderate .36, while reported correlations between CBCL Internalizing and Externalizing scales are frequently higher. For example, Achenbach and Rescorla (2000, pp. 159-160) report a correlation of .59 between standardized Internalizing and Externalizing scores on the CBCL/1.5-5 within a general population of preschoolers. Thus the strong correlation

between ITSEA Externalizing and CBCL/2-3 Internalizing may reflect limited differentiation of Internalizing and Externalizing on the CBCL, more than on the ITSEA.

- Because the PSI and the CCTI do not tap the same constructs as does the ITSEA, the extent to which reported associations reflect the validity of the ITSEA is less clear than with the CBCL/2-3. Nonetheless, reported results do suggest that parents who report problematic relationships with their young children, and who indicate that their children are temperamentally difficult (including high levels of emotionality, and low levels of soothability and sociability) are also more prone to report relatively higher levels of problem behaviors and less competence in their young children on the ITSEA.
- Analyses comparing parent ITSEA reports and developmental assessment results also support the authors' view that these assessments should demonstrate low associations with ITSEA scores reflecting problem behaviors, but higher associations with Competence. The moderate to high correlations of Competence with developmental assessment scores underscores the need to use age-normed measures (such as the ITSEA) for evaluating children's socioemotional adjustment.
- The validity of the ITSEA was also supported by findings of significant associations between ITSEA scales and assessments made by early intervention providers following 2-hour home visits. An examination of the pattern of these correlations, however, suggests that the strength of associations varied dramatically across different behavioral domains; agreement between parents and evaluators was considerably higher in domains involving very atypical behaviors, while parents and evaluators were less likely to agree strongly in domains reflecting less overt problems and competencies. It is likely that competencies and more subtle behavioral issues may not be as obvious during a relatively brief home visit as are very atypical behavioral problems. Additionally, although the evaluators were identified as early intervention providers, it is not clear what the background of evaluators was, with respect to training and experience.
- No information was provided regarding reliability or validity within narrower age bands, and no reliability or validity information was provided for children over 36 months of age. Additional work is needed in order to determine the extent to which reliability and validity vary across the ages covered by this measure.

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

Carter, Garrity-Roukous, Chazan-Cohen, Little, and Briggs-Gowan (2001) reported a series of analyses examining associations between maternal depression and parenting, infant-mother attachment, and toddler behavior problems and competencies. The sample included 69 mothers. Maternal depression was assessed with self-report and clinical interview measures at multiple timepoints, both prenatally and after the birth of the child. Also assessed were maternal emotional availability in a face-to-face play episode at 4 months of age, and infant-mother attachment at 14 months of age. When the toddlers were 30 months of age, mothers completed the ITSEA as well as the CBCL/2-3 (Achenbach, 1992).

- In one set of analyses comparing boys and girls whose mothers had no history of any psychopathological condition with boys and girls whose mothers had a lifetime history of depression (based on DSM-III-R criteria as assessed via a clinical interview), no

significant group differences in ITSEA Externalizing, Internalizing, Dysregulation (referred to as Regulation by Carter *et al.*, 2001), or Competence domain scores were found (see Carter *et al.*, p. 22).

- The authors also report bivariate correlations between ITSEA Externalizing, Internalizing, Regulation, and Competence scores and maternal depressive symptoms and emotional availability (see Carter *et al.*, 2001, p 23). Three depressive symptoms measures were used in these analyses – a prenatal measure, a postnatal measure composed of the mean of depressive symptoms assessed at infant ages 4 and 14 months, and a concurrent measure at toddler-age 30 months. At each age, depressive symptoms were self-reported using the Center for Epidemiologic Studies-Depression Scale (CES-D; Radloff, 1977).
    - ITSEA Externalizing was significantly correlated only with maternal emotional availability at infant-age 4 months (r = -.27). Mothers who were rated as more emotionally available in play interactions with their young infants also rated their children lower in Externalizing symptoms at 30 months of age. The correlation was not significant for either boys or girls separately, however.
    - Boys' ITSEA Internalizing scores were positively and significantly correlated with maternal concurrent depressive symptoms (r = .44). For girls, this correlation was negative (r = -.13), nonsignificant, and significantly lower than the correlation for boys. For girls, the only significant association was with maternal emotional availability—more emotionally available mothers rated their daughters lower in Internalizing symptoms (r = -.48). For boys, this association was nonsignificant (r = -.16); however, the difference between correlations for boys and girls was not significant. There were no significant correlations between ITSEA Internalizing scores and maternal characteristics for the full (i.e., boys and girls combined) sample.
    - ITSEA Dysregulation was significantly correlated with maternal prenatal depressive symptoms for the full sample (r = .26). Looking at correlations for boys and girls separately, it appears that this correlation was due entirely to a strong, significant correlation for boys (r = .52). Moderate significant correlations were also found between boys' ITSEA Dysregulation scores and maternal postnatal depressive symptoms and concurrent depressive symptoms (rs = .43 and .46, respectively). Mothers who were identified as exhibiting relatively high levels of depression at each assessment rated their toddler sons as having more regulation problems. In all cases, correlations for girls were negative, nonsignificant, and significantly lower than the correlations for boys (rs = -.10, -.21, and -.12 for prenatal, postnatal, and concurrent maternal depressive symptoms, respectively). Maternal emotional availability was not significantly associated with Dysregulation.
    - Significant correlations were found between ITSEA Competence scores and maternal prenatal depressive symptoms (r = -.24), and maternal depressive symptoms at 30 months (r = -.29). As with the Problem domain scales, however, it appears that these correlations are accounted for by moderate significant correlations for boys (rs = -.38 and -.49 for prenatal and concurrent depressive symptoms, respectively). None of these associations were significant for girls (rs = -.03 and .10, respectively); the correlation involving concurrent depression was

significantly lower for girls than for boys. Neither maternal postnatal depressive symptoms nor maternal emotional availability were associated with ITSEA Competence scores.

Comments

- Carter *et al.* (2001) found that ITSEA scores were associated with maternal characteristics, including a continuous measure of depressive symptoms and a measure of observed emotional availability; however, a measure of lifetime history of clinical depression was not associated with ITSEA measures. It is impressive that measures obtained much earlier in infancy and prenatally were predictive of subsequent ITSEA scores. Whether these associations are due to actual differences in children's behavior, however, rather than maternal perceptions, remains to be further investigated.
- The authors suggest that "Maternal depressive symptoms from pregnancy to 30 months postpartum appear to play a larger role in the emergence of problem behaviors for boys than for girls, while the quality of the early mother-infant interaction appears more central for girls' development" (Carter *et al.*, 2001, p. 23). The first part of this conclusion appears to be well supported by the data. The latter part, however, receives support from only one correlation presented in this report—the significant negative correlation between girls' Internalizing scores and maternal emotional availability. Associations between 14-month attachment classifications and ITSEA scale scores wee not reported. Although the reported findings involving sex differences in associations between ITSEA scores and maternal characteristics are intriguing, further studies with larger samples and multiple reporters of children's behavior problems and competencies will be important.

V. Adaptations of Measure

The Brief Infant Toddler Social and Emotional Assessment (BITSEA)

Description of Adaptation

The BITSEA (Briggs-Gowan, Carter, Irwin, Wachtel, & Cicchetti, 2004) was designed as a brief screener for socioemotional and behavioral problems. The BITSEA was developed and its psychometric properties were assessed using the same samples as were used for the ITSEA, and the 42 items included in the BITSEA were drawn from the pool of ITSEA items. Items were selected for inclusion based primarily on two criteria: 1) a majority of 12 mental health experts who served as consultants rated the item as being clinically important, and/or 2) the item had the highest factor loading on the ITSEA scale with which it was associated. Two items that did not meet these criteria were selected for inclusion; one item was deemed to be most representative of ITSEA Prosocial Peer, and one item was chosen because of clinical relevance. All items in the final BITSEA measure are found on the ITSEA as well, with two exceptions: Two items, one dealing with fearfulness and the second asking about sadness and depression, are composites of multiple ITSEA items. Three scales are derived from the BITSEA – Problem, Competence, and a combined Problem and/or Competence scale. Standardized scores are not provided for the BITSEA, however cut-points are provided within age bands for boys and girls separately, indicating scores on the Problem and Competence scales that might indicate that the child should receive further assessment.

According to Briggs-Gowan *et al.* (2004), the BITSEA can be completed in 5 to 7 minutes. A fourth to sixth grade reading level is required.

Psychometrics of Adaptation

- *Associations with ITSEA scores.* BITSEA Problem scores correlated strongly with ITSEA Internalizing, Externalizing, and Dysregulation domain scores (correlations of .58, .75, and .75, respectively), and had a low but significant negative correlation of -.20 with ITSEA Competence. BITSEA Competence scores correlated strongly with ITSEA Competence.
- *Internal consistencies* (alphas) for parent-report were .79 for Problem and .65 for Competence. The alpha for the combined Problem and/or Competence scale was not reported. Alphas for childcare provider reports were similar, .80 for Problem and .66 for Competence.
- *Test-retest reliability* was .87 for the Problem scale and .85 for the Competence scale with a 10 to 45 day interval. One-year stability of the BITSEA scales was .65 for Problem, and .53 for Competence.
- *Inter-rater agreement* was assessed using intraclass correlations across parents and between a parent and a childcare provider. Agreement between parents was .68 for the Problem scale and .61 for Competence. Parent-childcare provider agreement was .28 for the Problem scale and .59 for Competence.
- *Criterion-related validity* was assessed in two ways, both involving comparisons with BITSEA scores and scores on the Child Behavior Checklist for ages 1.5 to 5 years (CBCL/1.5-5; Achenbach & Rescorla, 2000).
    - First, associations between BITSEA scales and CBCL/1.5-5 scores were examined. Concurrent correlations between the BITSEA Problem scale and CBCL/1.5-5 Internalizing, Externalizing, and Total Problem scores were .64, .63, and .71, respectively. In contrast, correlations between BITSEA Competence scores and CBCL/1.5-5 Internalizing, Externalizing, and Total Problem scores were low to moderate, negative (-.23, -.31, and -.30, respectively), and significantly lower than the correlations with BITSEA Problem scores (see Briggs-Gowan *et al.*, 2004).
    - A second set of analyses examined the extent to which children whose CBCL/1.5-5 scores were within clinical or subclinical ranges (indicating the presence of possible behavioral and emotional problems) also had scores exceeding clinical cut-points on the BITSEA. These analyses included all children with CBCL/1.5-5 scores in the clinical or subclinical ranges, and an equal number of randomly selected children whose scores were within the normal range. In this sample, BITSEA Problem cutpoints correctly identified 93 percent of children whose CBCL/1.5-5 scores were within the clinical range, and 81 percent of children whose scores were in either the clinical or subclinical ranges. Further, 78 percent of children whose CBCL/1.5-5 scores were not in the clinical range, and 83 percent of children whose CBCL/1.5-5 scores were in neither subclinical nor clinical ranges, had BITSEA Problem scores that did not exceed clinical cutpoints (see Briggs-Gowan *et al.*, 2004).

- *Discriminant validity* was addressed by examining associations between BITSEA scores with scores on the MacArthur Communicative Development Inventory Short Form (MCDI-SF; e.g. Fenson, Pethick, Renda, Cox, Dale, & Reznick, 2000), a parent-report vocabulary checklist. Briggs-Gowan *et al.* (2004) reported that 23 percent of children with low BITSEA combined Problem and Competence scores had low scores on the MCDI. Looking at BITSEA Competence alone, 33 percent of those with low scores also had low scores on the MCDI. The authors interpreted these findings of low to moderate associations between the two measures as supportive of the discriminant validity of the BITSEA.
- *Predictive validity* was addressed by correlating BITSEA Problem and Competence scores with scores on the CBCL/1.5-5 and the ITSEA, administered one year later (see Briggs-Gowan, *et al.*, 2004).
    - Correlations between BITSEA Problem scores and CBCL/1.5-5 Internalizing, Externalizing, and Total scores were .45, .48, and .45, respectively; correlations with ITSEA Internalizing, Externalizing, Dysregulation, and Competence scores were .45, .57, .55, and -.24, respectively.
    - Correlations between BITSEA Competence scores and CBCL/1.5-5 Internalizing, Externalizing, and Total scores were -.12, -.20, and -.18, respectively; correlations with ITSEA Internalizing, Externalizing, Dysregulation, and Competence scores were -.11, -.21, -.13, and .63, respectively.

Studies Using Adaptation

No additional studies using the BITSEA were found.

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 the authors

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 U.S. population of preschool-age children. Participants were drawn 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.

III. 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.

IV. 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.

V. 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

No additional studies using this version were 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.

Social Skills Rating System (SSRS)

I. 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.

II. 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.

III. 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[18] (see Gresham, & Elliot, 1990, pp. 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,[19] 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 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.

[18] Gresham and Elliott include these correlations as evidence of convergent validity, rather than interrater reliability.
[19] 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.

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).[20]

In their study of low-income, rural white children, Pedersen *et al.* (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).

[20] 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.

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 *et al.* (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 *et al.* (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 *et al.* (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 *et al.* (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.

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

None found.

V. 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) was 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

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 *et al.* (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 *et al.* (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, pp. 94-95).

Concurrent and predictive validity
Sparrow *et al.* (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 *et al.* (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 *et al.* (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.

References for Social-Emotional Measures

Abidin, R. R. (1990). *Parenting Stress Index short form – test manual.* Charlottesville, VA: Pediatric Psychology Press.

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 & 1991 profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1992). *Manual for the Child Behavior Checklist/2-3 & 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*, 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*, 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*, 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. (1993). *Bayley Scales of Infant Development, Second Edition: Manual.* San Antonio, TX: The Psychological Corporation.

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., Irwin, J. R., Wachtel, K., and Cicchetti, D. V. (2004). The Brief Infant-Toddler Social and Emotional Assessment: Screening for social-emotional problems and delays in competence. *Journal of Pediatric Psychology. 29*, 143-155.

Briggs-Gowan, M. J., Carter, A. S., Skuban, E., & 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.

Burks, H. F. (1977). Burks' Behavior Rating Scales (BBRS). Austin, TX: PRO-ED.

Buss, A. H., & Plomin, R. (1975). *A temperament theory of personality development.* New York: Wiley Interscience.

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.

Carter, A. S., Briggs-Gowan, M. J., Jones, S. M., & Little, T. D. (2003). The Infant Toddler Social and Emotional Assessment (ITSEA): Factor structure, reliability, and validity. *Journal of Abnormal Child Psychology, 31,* 495-514.

Carter, A. S., & Briggs-Gowan, M. J. (2001). *Infant-Toddler Social and Emotional Assessment (ITSEA). Manual, Version 1.1.* Unpublished manual.

Carter, A. S., Garrity-Rokous, E., Chazan-Cohen, R., Little, C., & Briggs-Gowan, M. J. (2001). Maternal depression and comorbidity: Predicting early parenting, attachment security, and toddler social-emotional problems and competencies. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 18-26.

Cohen, N. J. (1983). Mother-child interaction in hyperactive and normal kindergarten-aged children and the effect of treatment. *Child Psychiatry & Human Development, 13*, 213-224.

Conners, C. K. (1989a). *Conners' Teacher Rating Scales*. North Tonawanda, NY: Multi-Health Systems.

Conners, C. K. (1989b). *Conners' Parent Rating Scales*. North Tonawanda, NY: Multi-Health Systems.

Conners, C. K. (1990). *Manual for Conners' Rating Scales*. Toronto: Multi-Health Systems.

Conners, C. K. (1995). *Conners' Continuous Performance Test (CPT)*. Toronto: Multi Health Systems.

Conners, C. K. (1997). *Conners' Rating Scales – Revised: Technical manual*. North Tonawanda, NY: Multi-Health Systems.

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, 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*, 255-269.

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

Elliott, S. N., Barnard, J., & Gresham, F. M. (1989). Preschoolers' social behavior: Teachers' and parents' assessments. *Journal of Psychoeducational Assessment, 7*, 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*, 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*, 199-214.

Fenson, L., Pethick, S., Renda, C., Cox, J., Dale, P., & Reznick, J. (2000). Short-form versions of the MacArthur Communicative Development Inventories. *Applied Psycholinguistics, 21*, 95-115.

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*, 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*, 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,* 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*, 1-6.

Gresham, F. M., & Elliott, S. N. (1990). Social Skills Rating System: Manual. *Circle Pines, MN: American Guidance Service.*

Hans, S. L., & Jeremy, R. J. (2001). Postneonatal mental and motor development of infants exposed in utero to opioid drugs. *Infant Mental Health Journal, 22*, 300-315.

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*, 103-111.

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*, 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*, 26-41.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children*. 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*, 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.

Lachar, 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*, 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*, 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*, 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*, 201-219.

LeBuffe, P. A., & Naglieri, J. A. (1999). *Devereux Early Childhood Assessment Program: Technical manual.* Lewisville, NC: Kaplan Press.

McCarthy, D. (1972). *McCarthy Scales of Children's Abilities*. New York: Psychological Corporation.

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*, 358-375.

Matheny, A. P., & Wilson, R. S. (Nov. 1981). Developmental tasks and rating scales for the laboratory assessment of infant temperament. *Catalog of Selected Documents in Psychology. 11 MS. 2367*, 81-82.

Merydith, S. P. (2001). Temporal stability and convergent validity of the Behavior Assessment System for Children. *Journal of School Psychology, 39*, 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.

Mullen, E. M. (1989). *Infant Mullen Scales of Early Learning.* Circle Pines, MN: American Guidance Service.

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.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401.

Reynolds, C. R., & Kamphaus, R. W. (1998). *BASC Behavioral Assessment System for Children: Manual.* Circle Pines, MN: American Guidance Service.

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.* Available: 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*, 1760-1767.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland Adaptive Behavior Scales.* Circle Pines, MN: American Guidance Service.

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

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.

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

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, 637-643.

Vig, S., & Jedrysek, E. (1995). Adaptive behavior of young urban children with developmental disabilities. *Mental Retardation, 33*, 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*, 151-161.

Werner, E.E., & Smith, R.S. (1982). *Vulnerable but invincible: A longitudinal study of resilient children*. New York: Adams, Bannister, Cox.

Wolf, A. W., & Lozoff, B. (1985). A clinically interpretable method for analyzing the Bayley Infant Behavior Record. *Journal of Pediatric Psychology. 10*, 199-214.

Early Head Start Measures

Early Head Start – List of Measures

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.

I. 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)

II. 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

III. Mastery

- Dimensions of Mastery Questionnaire (Morgan et al, 1990)

IV. 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

V. Parenting

- Child Abuse Potential Inventory
- Parenting Stress Inventory
- Stony Brook Family Reading Survey

VI. 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

VII. Quality of the Home Environment

- Home Observation for Measurement of the Environment (HOME) *

VIII. 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

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).

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

Comments
- 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 (α = .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).

Comments

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 ($\alpha = .66$ for both scales) were used in Early Head Start impact analyses.

IV. 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.

V. Adaptations of Measure

The Early Head Start Research and Evaluation Project

- 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* The Early Head Start Research and Evaluation Project

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.

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
Time
8 to 9 minutes.

Cost
Estimated cost of graduate student training and videotaped coding is $95 per videotape

Comments
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)

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

None.

Comments

- 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

V. Adaptations of Measure

None.

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 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

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

Time

12 to 13 minutes.

Cost

Estimated cost of graduate student training and videotape coding is $95.00 per videotape.

Comments
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
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. 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%

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)

IV. 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).

Comments

- 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

V. Adaptations of Measure

None found.