The report provides detailed information about the methods and instruments used to evaluate school readiness initiatives, discusses important considerations in selecting instruments, and provides resources and recommendations that may be helpful to those who are designing and implementing school readiness evaluations.

OVERVIEW

This review of evaluations of the effectiveness of publicly funded state and local school readiness initiatives describes the instruments used to measure child outcomes and features of the instruments themselves. The report is organized around three questions:

- How did evaluations of state- and locally funded school readiness programs collect data on child outcomes? (What types of evaluations were conducted? When were data collected? What domains of children's school readiness were measured—cognitive, social, behavioral adjustment?)
- What instruments or measures were used to collect child outcome data in evaluations of state-funded school readiness initiatives?
- What are the key features of these instruments (developer, administration, purpose, age group, psychometric properties)?

The results indicate that state and local evaluators have used a variety of instruments to collect child outcome data, some that are well known and others that are not. In general, many of the well known instruments demonstrate adequate psychometric properties (reliability and validity, which ensure that the instruments consistently measure what they were intended to measure), but a number of issues, such as the appropriateness of the measure to the study's purpose and sample, appear to present substantial challenges in evaluations of state- and locally funded school readiness programs.

Selecting and implementing instruments for evaluating school readiness programs are not easy. The findings of this report highlight the challenges that evaluators face in ensuring that data are collected in a manner that yields credible, trustworthy, and meaningful information about child outcomes. The report offers several recommendations based on the data collected from this sample of school readiness evaluations to help school readiness programs and evaluators as they select instruments for assessing school readiness programs and implementing the evaluations. The

report also lists a number of useful resources to assist evaluators in making decisions about child assessments: resources to guide decisions about how to assess child outcomes, reviews of measures, and web sites with technical information related to measures used in large federal studies.

WHY THE STUDY IS NEEDED

The link between high-quality early education programs and improved student outcomes is fairly well established. High-quality early education can promote positive developmental outcomes both while children are enrolled and as they enter school. Studies (Barnett, 1995; Guralnick, 1997; Karoly, Greenwood, Everingham, Hoube, Kilburn, Rydell, Sanders, & Chiesa, 1998) have found that children who receive high-quality early education perform better on measures of cognitive and language development and are less likely to be held back later in school. The research also indicates that young children living in situations that place them at greater risk of school failure, such as poverty and low levels of maternal education, benefit most from quality early childhood services (Brown & Scott-Little, 2003).

As the number of new school readiness programs has increased, so has the pressure to provide data on program effectiveness

In response, states and local communities have created a wide variety of school readiness programs and initiatives, particularly for children identified as "at risk" of school failure. Head Start and Even Start are examples of federal school readiness programs. Examples of state programs are North Carolina's Smart Start program, Michigan's prekindergarten program, and Georgia's lottery-funded school readiness program. According to recent statistics (Pre-K Now, 2006), 41 states plus the District of Columbia offer their own classroom-based prekindergarten programs for at least some of their three- to five-year-old children, up from 10 states in 1980.

Local school districts are also becoming involved. A survey of 16,000 school districts across the nation revealed that 17 percent of districts receiving Title I funds used a portion of the funds for preschool services during the 1999/2000 school year (McCallion, 2004), reaching about 8 percent of children who will eventually enter public kindergarten.

As the number of new programs has increased, so has the pressure to provide data on program effectiveness. Policymakers, educators, and parents want to know what benefits children receive from participating in these programs, and they want to know whether the programs prepare children for success in school. As policymakers have provided increased resources for early childhood programs, they have put in place new requirements for collecting data to demonstrate the effects of the program. As a result, there has been a dramatic increase in demands that early childhood programs present evidence of child outcomes and an emphasis on program accountability. Examples of state efforts to collect data on how successful early childhood initiatives have been in preparing children for success in school include:

- California's Desired Results for Children and Families system, which is designed to help educators document the progress children and families make and to provide information to help practitioners improve their services (http://www.cde.ca.gov/sp/cd/ci/desiredresults.asp).

- Florida's Voluntary Prekindergarten Program, which requires an analysis of how children who have completed the program perform on a readiness assessment and includes sanctions for programs in which children do not score at the acceptable rate within two years. Legislation requires that local school readiness coalitions collect pre- and post-test data on children's readiness for school (Florida Statutes, 2006).

- The Maryland Model for School Readiness, which is an assessment and instructional

system designed to gauge children's skills and knowledge at kindergarten entry (http://www.mdk12.org/instruction/ensure/MMSR/).

Although much has been written about the effectiveness of small model early childhood programs (such as the Abecedarian and Perry Preschool Project), until recently there has been little evidence to document whether large-scale, publicly funded programs, such as those in the examples above, actually help children perform better in school. Recent meta-analyses of the effectiveness of such programs found that the large-scale publicly funded school readiness interventions studied had moderate effects on a variety of child outcomes (Brown & Scott-Little, 2003; Gilliam & Zigler, 2001; Gilliam & Zigler, 2004).

This evidence on program effectiveness has been criticized for the design and quality of the evaluations. Gilliam & Zigler (2001, 2004) maintain that there were serious methodological weaknesses in many of the evaluations and that some were so flawed that it was difficult to draw any clear conclusions about program effects. Most of these studies relied primarily on a one-group, pre-test–post-test design. A large number relied on checklists, teacher ratings, or teacher-survey results as the outcome measures of focus.

Similar findings were reported by Brown & Scott-Little (2003) in a review of school readiness evaluations. They noted the importance of using consistent, reliable, and valid measures to document the effectiveness of the programs. Although some of the studies used well known, standardized instruments, a number of studies used instruments designed specifically for the study and provided little or no reliability or validity data (see box 1 on key terms).

Given the increasing number of school readiness evaluations that have been published and the questions raised by researchers who have reviewed the evaluations, there is a need to better understand how child outcomes have been measured in evaluations. This report provides data on the instruments that have been used in evaluations. Furthermore, this report addresses questions and concerns raised by decisionmakers who are selecting assessment instruments and designing evaluations of state and local school readiness programs. Decisionmakers can look to this report for information on how programs across the country have evaluated child outcomes and as a resource for basic information about commonly used assessment instruments. The report examines evaluations of the effectiveness of publicly funded state and local school readiness initiatives and describes the range of methods and instruments used to collect child outcome data and evaluate the initiatives, summarizing information on the technical properties of the instruments used. The review highlights commonly used instruments and the challenges evaluators face when collecting child outcome data.

There is a need to better understand how child outcomes have been measured in evaluations

While the report supplies important information on how state- and local-level school readiness evaluations have collected child outcome data, there are several limitations or alternate purposes that the report was not designed to address. It is not a detailed guide for how to develop and deploy an effective evaluation effort or a comprehensive comparison of instruments.

Many factors must be considered when developing and implementing program evaluations, and selection of appropriate instruments is but one piece of the overall plan. (The final section of the report suggests additional resources that may be helpful in formulating an evaluation plan and identifying specific instruments.) Furthermore, the report's coverage of instruments that may be used by programs in assessing young children is not exhaustive. The report identifies only the instruments that were used to evaluate state- and locally funded school readiness initiatives that were located during the research search process (see box 2). Federal government funded large-scale

BOX 1

Key terms in the report

Issues related to assessing young children and evaluating school readiness programs are complex and technically challenging. The following key terms are important for understanding the analysis presented here. The definitions are compiled from a number of sources, including the web site of the Council of Chief State School Officers (http://www.ccsso.org); Cohen & Spenciner (2007); Epstein, Schweinhart, Debruin-Parecki, & Robin (2004); and Trochim (2006).

Criterion-referenced assessment. An instrument that compares an individual's performance to a predetermined criterion or standard to determine whether the individual has met the standard.

Norm. The scores obtained by a standardization sample to which examinees' test scores are compared.

Norming group or sample. A sample used to standardize a test. The norming sample should consist of a sufficiently large random sample taken from the target population.

Norm-referenced assessment. An assessment that compares an individual's performance with the performance of a norming group or sample.

Psychometric properties. Characteristics of instruments such as reliability and validity that ensure that an instrument consistently measures what it was intended to measure.

Reliability. Reliability refers to the consistency with which a test yields similar results over time, even when administered in different forms or by different examiners. Several different forms of reliability may be used to estimate the stability of test scores for a group of examinees. *Internal consistency* focuses on the degree to which individual test items are related. *Test-retest* refers to the consistency of an instrument from one time to the next. A test-retest reliability coefficient is obtained by administering the same test twice to the same group and correlating the scores. *Split-half* divides test items in two based on randomly selecting items (odd versus even, or other rules). The split-half reliability estimate is the correlation between the two total scores. *Alternate form* reliability is computed by correlating the scores on pairs of alternate forms of a test designed to be administered interchangeably to the same examinees. *Interrater* refers to the degree to which different raters or observers give consistent estimates of the same phenomenon.

Standardized assessment. A testing instrument that is administered, scored, and interpreted in a standard manner. It may be either norm referenced or criterion referenced.

Validity. Validity refers to the degree to which an instrument measures what it is supposed to measure. No single type of validity is appropriate across all instruments. The type of validity that is most important varies with the purpose for which the data are collected. *Content validity* refers to the extent to which the components of an instrument include the relevant content domain for the construct. This type of validity assumes that there is a good detailed description of the content domain. Content validity is typically established through the professional judgment of experts, who assess the extent to which the assessment items adequately address the constructs the assessment is seeking to measure. *Criterion validity* refers to the degree to which an instrument correlates with other measures of the same construct assessed either concurrently or predictively. Thus *concurrent criterion validity* refers to the extent to which a test correlates with an external criterion measured at the same time, and *predictive criterion validity* refers to an instrument's ability to predict an individual's performance in specific abilities.

early childhood studies are not covered. While these studies have substantially expanded the knowledge base and led to the development of new instruments, they are not the focus of this report, which addresses the questions often raised by state- and local-level evaluators about how other state and local programs have collected child outcome data.

DESCRIPTION OF THE EVALUATION REPORTS REVIEWED

The sample for this review includes 82 evaluation documents (see box 3 and appendixes A and B on the search process and coding of documents and appendix C for the list of evaluation reports) covering 78 studies, where a study is

BOX 2

Key features of the methods and instruments reported on in the results section

This box summarizes some of the features of the instruments that are coded and reported on in the results section.

What child outcomes were assessed? School readiness is a multifaceted construct. Experts consider a number of skills and abilities to be important for success in school, ranging from academic-related skills to social skills, emotional readiness, physical abilities, and attitude toward learning tasks. The review looked at what elements of children's development and learning were assessed in program evaluations.

When were child outcomes measured? Program evaluators must decide when to collect data—at the end of the school readiness program, when children start kindergarten, or later in the children's school career.

How were child outcomes measured? There are a variety of ways to collect data on children's development and learning:

- Direct assessments, in which the examiner asks children to perform specific tasks.
- Checklists and rating systems that ask teachers or parents familiar with a child to indicate how well a child can do specific tasks or knows certain content.
- Naturalistic observations, in which a teacher or other observer records children's activities in their regular classroom setting.
- Achievement tests, typically given in later elementary grades.
- Record reviews, to determine what services children have received during their education.

What was the purpose of the instrument? Assessment instruments are designed for specific purposes, and it is commonly recommended that they be used only for the purpose for which they were designed. In general, these purposes can be described within the following categories (Epstein and others 2004; National Education Goals Panel, 1998; Scott-Little & Niemeyer, 2002):

- *Screening.* Brief child assessment can be used to identify children with suspected disabilities or who are at risk of failing in school. Further evaluation is required to pinpoint a potential learning problem or a need for specialized services.
- *Planning instruction.* Assessments may be used to support instruction by providing information on children's strengths and weaknesses that teachers can use to plan activities. Instructional assessments are usually conducted over time to provide information on a child's skills and abilities in a variety of learning situations and are often closely related to the curriculum.
- *Identifying program improvement needs.* Child assessments can identify program areas in need of improvement, such as curriculum, resources, and materials. Combined with implementation data, these types of child assessments can provide formative data that can benefit the program and staff.
- *Monitoring changes over time and determining effectiveness.* This type of assessment provides information on children's skills and abilities to track growth on important outcomes over time and to provide data to evaluate programs.

Was the instrument appropriate to the age of the children assessed? Assessments are targeted to the skills, abilities, and knowledge of children within a specific age range (Scott-Little & Niemeyer, 2001).

Have the instruments demonstrated adequate reliability and validity? The usefulness of evaluation results is determined, in part, by the quality of the instruments used to collect the data. Reliability and validity are important indicators of an instrument's technical properties.

Unreliable instruments may overestimate or underestimate child outcomes in unpredictable ways. Reliability is typically reported as correlation coefficients (in the range 0–1.0), and the higher the coefficient, the more

(CONTINUED)

BOX 2 (CONTINUED)

Key features of the methods and instruments reported on in the results section

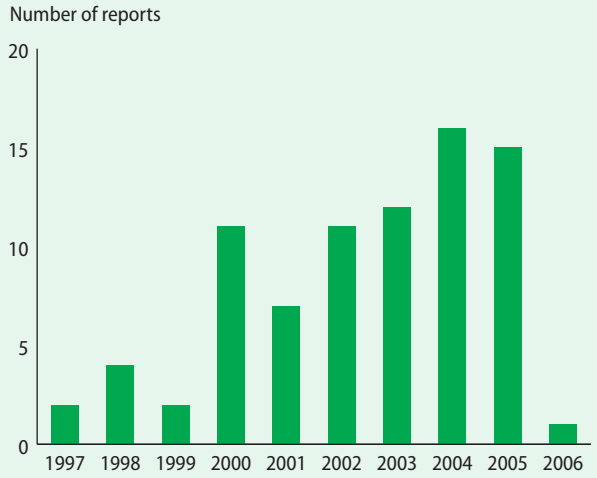
reliable the instrument (Kisker and others 2003). For making educational or placement decisions about individual children, instruments should have reliability coefficients of .90 or higher, but for assessing outcomes across groups of children, instruments with reliability coefficients of .80 and higher are commonly accepted as adequate (Cohen & Spenciner, 2007).

Validity—the extent to which an instrument measures what it is supposed to measure and is appropriate to the purpose for which it is being used—must also be considered (Kisker et al., 2003). Three forms of validity are particularly important in school readiness evaluations—content, concurrent, and predictive validity (see box 1). Like reliability, concurrent and predictive validity are often reported as a correlation coefficient, with higher values indicating stronger relationships between the results of the instruments being compared.

A related issue concerns the norming group or sample. The use of norm-referenced instruments allows evaluators to make statements about how well the children who were assessed did compared with other groups of children who have taken the same test. If the children who are assessed are different from the norming group on whom the norms are based, however, it may not be valid to compare their scores to the norms that are provided for the instrument (Cohen & Spenciner, 2007).

FIGURE 1

The number of evaluation reports in the review sample rose dramatically after 2000



Source: Authors' analysis based on data from evaluation reports reviewed; see box 3 and appendixes A and C.

defined by the specific population sample. For example, reports that follow a cohort of children over time (such as the Georgia Prekindergarten Program and Washington's Early Childhood Education and Assistance Program) were coded as a single study. A program that used the same methodology and dependent variables but added new cohorts each year was coded as multiple studies.

The sample includes reports published between 1997 and the first half of 2006. Before 2000 only a handful of programs had published evaluation reports designed to document program improvement and accountability (figure 1). Since then the number has increased dramatically, perhaps reflecting an increasing emphasis on documenting program outcomes or on holding programs accountable.

DESCRIPTION OF EVALUATED PROGRAMS

The sample of reports examined in this analysis included a mix of state- and locally funded programs using a variety of service-delivery models. The 82 reports covered 41 separate programs, 26 of them state-funded readiness initiatives. Most are administered by the state department of education or another state agency, such as Georgia's Department of Early Care and Learning and North Carolina's Partnership for Children. The remaining programs are local programs supported by other funding, such as federal Title I funding. Table 1 lists the state-funded programs represented among the evaluation reports and the reports for each program. Table 2 lists the same information for the locally funded programs.

Within the 41 programs children were served in school-based and community-based settings. At

BOX 3

Document search and coding process

Between May and July 2006 an extensive search was conducted of state department of education web sites for information and evaluation reports related to each state's school readiness initiatives. In addition, early childhood specialists were contacted and asked whether they had any evaluation reports that might document the effectiveness of their state's prekindergarten programs. Information was also sought on the web sites of such organizations as Child Trends, National Center for Early Development and Learning, Mathematica Policy Research, RAND Corporation, Southern Regional Education Board, U.S. Department of Health and Human Services, and the regional educational laboratories. Finally, conference programs and proceedings were reviewed for the American Educational Research Association, Head Start National Research Conference, the National Association of Elementary School Principals, and the Society for Research in Child Development.

The search uncovered 217 articles, reports, conference presentations, and dissertations that had been disseminated between January 1997 and July 2006. The titles and abstracts identified through the search were next screened for relevance, resulting in the identification of 148 documents as candidates for inclusion in the study. These documents were then subjected to the following screening criteria, to establish whether they:

- Evaluated school readiness initiatives in the United States.
- Evaluated publicly funded programs or interventions that target children from birth to age five to enhance their readiness for school and that include a component of classroom-based services.
- Presented some type of child outcome data indicating the effectiveness of the program.
- Provided sufficient information for the study to be coded.

Of the 148 documents, 82 met the inclusion criteria and were coded to capture a general description of the school readiness program and the methods used to evaluate its effectiveness (see appendix B). The coding identified each instrument in the evaluation studies.

Next, information was collected about each instrument using a separate coding process. This coding process was developed to capture a general description of the assessment, its purposes and target population, summary information about how each instrument is administered, and the scores and scoring procedures and publicly available technical information for each instrument (validity and reliability). Information for the coding process was taken from materials that accompany each assessment instrument and from each publisher's web site. A summary of the information collected about the more commonly used assessment instruments is available in appendix D.

least 11 programs served children in classrooms, primarily in school settings, while 6 programs served children primarily in community-based sites. Nineteen programs served children in mixed settings, with classrooms established in both school-based and community-based sites. The remaining 5 programs were coded as "unspecified," because the information in the report was insufficient to determine the setting.

RESULTS FOR THE THREE RESEARCH QUESTIONS

Analysis of the evaluation reports included in this review shows how outcome data were collected in school readiness evaluations. The analysis includes answers to the three research questions, describing how the school readiness evaluations were conducted, what instruments were used, and what were their key features, based on publicly available information about 27 commonly used instruments.

Research question 1: How did evaluations of state- and locally funded school readiness programs collect data on child outcomes?

The coded data about how the school readiness evaluations described in the sample of reports were implemented include information on the type of evaluation conducted, the type of school

TABLE 1
State-level programs included in the evaluation report sample and reports by program

Program name	Citation
Arizona At-Risk Preschool Program	Norton (1997)
California, First 5 California	California Children & Families Commission (2003, 2004, 2005)
Connecticut School Readiness Program	Bond (2000)
Delaware Early Childhood Programs	McCormick-Gamel & Amsden (2002)
Georgia Universal Prekindergarten Program	Andrew Young School of Policy Studies (2000) Henderson, Basile & Henry (1999) Henry et al. (2001, 2003, 2004) Henry et al. (2005)
Georgia Summer Readiness Pilot Program	Ponder, Rickman, & Henry (2004)
Illinois Prekindergarten Program for At-Risk Children	Illinois State Board of Education (2001)
Iowa Shared Visions Programs for At-Risk Four-Year-Olds	Zan & Edmiaston (2000)
Kansas Four-Year-Old At-Risk Program	Martinez (2000, 2002)
Kentucky Preschool Program	Hemmeter (2001)
Louisiana LA4 Program	Louisiana Department of Education (2002, 2003)
Louisiana LA4 and Starting Points Prekindergarten Programs	Louisiana Department of Education (2004, 2005)
Louisiana Starting Points Program	Louisiana Department of Education (2001)
Michigan Full-Day Preschool Program	Jurkiewick et al. (2004)
Michigan School Readiness Program	Lamy, Barnett, & Kwanghee (2005) Xiang & Schweinhart (2002) Xiang et al. (2000)
Minnesota School Readiness 1999/2000	Mueller (2001)
Missouri HB 1519 Early Childhood Project	Thornburg, Fuger, Mayfield, & Mathews (2003)
Nebraska Early Childhood Education Grant Program	Jackson & St. Clair (2004)
Nevada Early Childhood Education Program	Leitner (2003)
New Jersey Abbott Preschool Program	Frede et al. (2004) Lamy et al. (2004, 2005a, 2005b)
North Carolina More at Four	Peisner-Feinberg & Maris (2003, 2005) Peisner-Feinberg, Elander, & Maris (2006)
North Carolina Smart Start	Bryant et al., (1998, 2003) Maxwell, Bryant, & Miller-Johnson (1999)
Ohio Head Start	Cogswell, Lochtefeld, Skaggs, & Walker (1998)
Oklahoma Prekindergarten Program	Gromley & Gayer (2003) Gromley et al. (2004, 2005)
South Carolina Child Development Program for Four-Year-Olds	Lamy et al. (2005) South Carolina Department of Education (2002, 2004) Yao, Snyder, Burnett, Lindsay, & Tenenbaum (2000)
Washington Early Childhood Education Assistance Program	NWREL (1998, 2000)
West Virginia Early Education Program	Lamy et al. (2005)
Total number of studies	52

readiness outcomes examined, the period the outcome data covered, and the type of data collectors and any training they received. The results are reported based on the number of studies rather than the number of reports.

Type of evaluations conducted. The primary question here was whether the evaluations focused on the child outcomes alone (summative evaluation) or whether they also collected data on the nature of the program (implementation). Thirty-one studies

TABLE 2
Local programs included in the evaluation report sample and reports by program

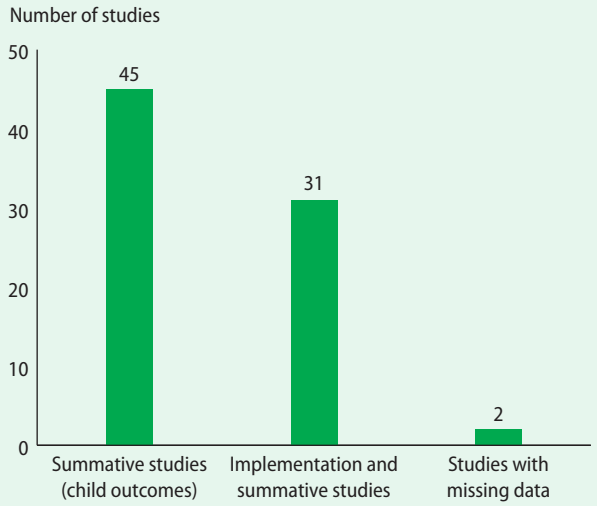
Program name	Citation
Allegany County, Maryland, Judy Center	eQuotient (2002)
Austin, Texas, Independent School District Prekindergarten Expansion Grant Program	Curry (2000, 2001, 2002, 2003, 2004, 2005)
Charlotte-Mecklenburg, North Carolina, Bright Beginnings	Smith et al. (2003)
Chicago, Illinois, Child-Parent Centers	Clements, Reynolds, & Hickey (2004) Reynolds (1997) Reynolds, Temple, Robertson, & Mann (2001) Temple, Reynolds, & Meidel (2000)
Columbiana County, Ohio, Head Start	Columbiana County Head Start (2004)
Dallas, Texas, Independent School District Prekindergarten Expansion Grant Program	Dallas Independent School District (2004, 2005)
Dallas, Texas, Language Enrichment Activities Program	Dallas Independent School District (2005)
District of Columbia Public Preschool Program	Marcon (2000)
Ft. Worth, Texas, Child Care Associates	Mindel (2005)
Garrett County, Maryland, Overlook Judy Center	Overlook Judy Center Partnership (2005)
Greenville, South Carolina, 4K Early Childhood Program	Coleman & McCreary (2003)
Los Angeles, California, Unified School District, School Readiness Language Development Program	Maddahian (1998)
Miami-Dade County, Florida, Prekindergarten Program	Levitt (2002)
Pittsburgh, Pennsylvania, Early Childhood Initiative	Bagnato (2002) Bagnato, Suen, Brickley, Smith-Jones, & Dettore (2002)
Rochester, New York, Early Childhood Assessment Partnership	Gramiak et al. (2004) Gramiak, Hightower, & Baker (2005) Montes et al. (2002) Montes et al. (2003) Montes & Hoffman (2004)
Saginaw, Michigan, Public Preschool Program	Kurecka & Klaus (2000)
Total number of studies	30

collected both summative and implementation data, 45 collected and reported only on summative data, and two were missing data (figure 2).

Types of school readiness outcomes evaluated. Program leaders and evaluators have to make decisions about which outcomes to study in assessing program impact. To see what type of school readiness outcomes were evaluated in the studies under review, child outcomes were categorized into three major areas: child development, school performance, and long-term outcomes. These areas were then subdivided. Child development includes eight domains: overall child development, cognitive development, general knowledge and awareness, language/communication skills, literacy skills, mathematics/pre-mathematics skills, child health/physical development, and social/emotional development. School performance focuses on six areas: school grades/report cards, standardized achievement test results, grade retention, special services (such as special education referral and speech/language services), attendance, and school adjustment/attitude. Long-term outcomes include three domains: dropout rate, delinquency, and arrest record. In most instances studies evaluated more than one area.

Type of child development outcomes examined. The most commonly assessed child development

FIGURE 2
Most evaluations looked at child outcomes (*N* = 78)



Source: Authors' analysis based on data from evaluation reports reviewed; see box 3 and appendixes A and C.

outcomes were literacy/pre-literacy skills, followed by language/communication skills, mathematics/pre-mathematics skills, social/emotional development, and child health/physical development.

The program studies most commonly included measures of children's *literacy skills* (50). These included children's interest in books, their knowledge and understanding of print, and their ability to recognize letters and identify sounds. Related to children's literacy skills are *language/communication skills.* These measures focus primarily on children's ability to communicate effectively—their ability to comprehend and express their understanding. At least 28 studies specifically assessed children's language and communication skills separately from literacy.

At least 41 studies reported collecting data on children's *mathematics/pre-mathematics skills.* Typically, measures of mathematics/pre-mathematics skills collected data on children's ability to understand numbers and shapes, solve mathematical operations, and identify solutions to problems.

Thirty-six of the studies reviewed reported that they collected information on children's *social/emotional development.* This includes such behaviors as children's social competence and pro-social behavior, as well as children's negative or problem behaviors. Pro-social behavior refers to behaviors that facilitate interaction, such as sharing, turn-taking, and cooperating with others. Problem behaviors are those that interfere with good social relationships, such as fighting and arguing.

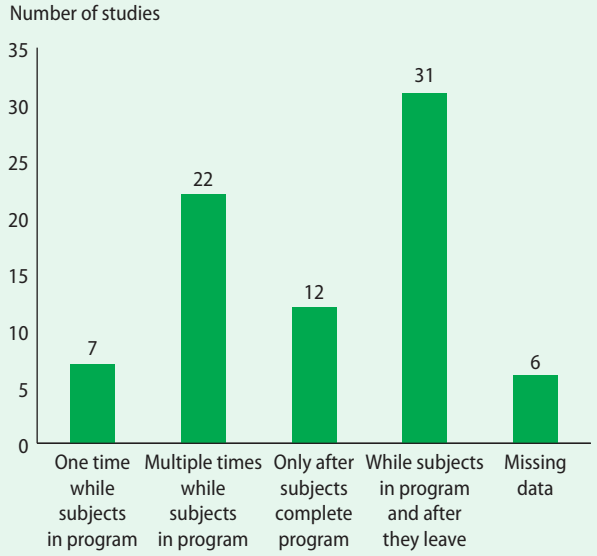
Finally, 30 studies focused on children's *health/physical development.* This domain includes measures of children's health and well-being; fine and gross motor movement; and nutrition, dental, and medical needs.

Type of school performance outcomes examined. Many studies continued to follow students once they completed the program and entered school. School outcomes were measured in a variety of ways, including report cards and grades (10), attendance (6), grade retention (11), school adjustment/attitude (9), special services (10), and standardized achievement tests (16). At least one study examined the number of times that children were expelled from school as an outcome measure.

Type of long-term outcomes examined. Only one study, Reynolds et al. (2001), followed children beyond the formal school years. The most recent evaluation focused on the long-term effects of this program—school dropout/high school completion and criminal arrests.

Data collection time frame. Data on child outcomes can be collected while children are enrolled in the program, at the end of their enrollment, or as they progress through their school careers. Most of the studies took a longitudinal approach, looking at outcomes when the children were enrolled in the school readiness program and continuing to collect data after the children left the program (31 studies; figure 3). The next most common approach was to collect data several times from children while they were enrolled in the program, but not to collect follow-up data after children entered school (22). Others assessed selected children once they entered kindergarten

FIGURE 3
Most of the studies collected data more than once (*N* = 78)



Source: Authors' analysis based on data from evaluation reports reviewed; see box 3 and appendixes A and C.

(12), when a comparison group became easier to recruit and follow (for instance, Michigan School Readiness Program). Several studies have evaluated multiple cohorts of children. Louisiana has evaluated five cohorts, the largest number.

Data collectors used in the studies. The types of people selected to administer the instrument or collect the data varied. A standardized test is typically administered by an examiner who has received specialized training. An informal assessment is more likely administered by a child's teacher or another professional who has frequent contact with the child (Epstein and others 2004). In almost all of the studies teachers were asked to administer or provide information on at least one of the instruments. Of the 68 instruments administered in the studies reviewed, 45 were administered by teachers. Outside evaluators were also used to collect data on children's performance. They were the sole examiners or data collectors for eight of the instruments (such as for Get Ready to Read[1]). In a few instances, other program staff, such as the director, provided information, but this occurred in only two of the studies.

The reports stated that training was provided in almost half (39) of the studies that were reviewed, but few studies provided details on the length or nature of the training. Among studies that provided additional description, only four reported that examiners were trained to adequate reliability, but none offered further information about the training. Only one program, the Rochester Early Childhood Assessment Partnership (Gramiak et al., 2004; Gramiak, Hightower, & Baker, 2005; Montes et al., 2002; Montes et al., 2003; Montes & Hoffman, 2004), reported the amount of training (three hours of training on the Child Observation Record[2] and the Teacher-Child Rating Scale[3]).

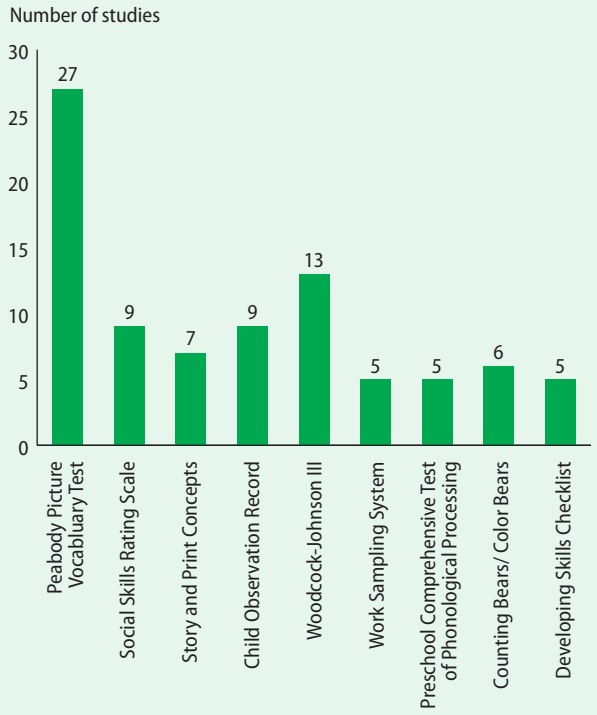
Research question 2: What instruments or measures were used to collect child outcome data?

Key to designing an effective evaluation is selecting appropriate instruments and measures. Researchers and program evaluators must consider the methods and instruments used to monitor change in young children. With rising demands for accountability both the assessment instrument and those responsible for collecting the data bear an increasing responsibility to accurately and fairly capture the outcomes associated with a program.

The studies used a variety of methods to assess child outcomes, including observational measures, standardized measures, checklists, teacher ratings, and teacher surveys. Most studies used more than one instrument to collect data on children's skills and abilities in more than one domain. A total of 68 instruments were used, with a mean number of 2.53 instruments and a range of 0 (in at least one of the studies for the Chicago Child-Parent Centers, data were obtained from a record review only [Reynolds et al., 2001]) to 13 (Georgia early childhood study [Henry et al., 2004; 2005]).

The Peabody Picture Vocabulary Test–Third Edition[4] was used most often, followed by the Woodcock-Johnson III,[5, 6] the Child Observation Record, and the Social Skills Rating Scale[7] (figure 4). Several instruments used in the Head Start

FIGURE 4
Most frequently used instruments



Source: Authors' analysis based on data from evaluation reports reviewed; see box 3 and appendixes A and C.

Family and Child Experiences Survey (FACES) are also being used to evaluate state- and locally funded school readiness programs (such as Color Bears and Counting Bears[8] and Story and Print Concepts[9]). Instruments that were used less frequently include the Work Sampling System,[10] the Preschool Comprehensive Test of Phonological Processing,[11] and the Developing Skills Checklist.[12]

Most of these are well known, psychometrically validated instruments. Some program evaluations, however, used locally developed instruments with little data on validity or reliability. Typically, these instruments were measures of teachers' perceptions about some dimension of children's development, such as social and emotional development or cognitive and language abilities. In five studies evaluators used study-specific instruments that asked teachers to rate children's readiness and likelihood of progressing to the next grade level. Evaluations of three of the programs (First 5 California, Georgia Universal Prekindergarten, and Washington Early Childhood Education Assistance Program) developed instruments that were designed to capture parents' perceptions of their children's adjustment and readiness for school.

This analysis also looked at whether programs tested the validity and reliability of their outcome measures, using the data from their own studies. Only four program evaluations—First Five California, the Michigan School Readiness Program, the Pittsburgh Early Childhood Initiative, and the Rochester Early Childhood Assessment Partnership—examined the reliability of their measures. Each study reported reliability coefficients of .85 or higher. Only two programs—the Charlotte-Mecklenburg, North Carolina Bright Beginnings and the Michigan School Readiness Program—reported validity data in their study reports.

In general, programs used instruments in their original format. Only two programs (Charlotte-Mecklenburg, North Carolina, Bright Beginnings and Pittsburgh, Pennsylvania, Early Childhood Initiative) reported that instruments were modified or adapted.

Research question 3: What are the key features of these instruments?

To gain a sense of the general properties of the instruments, 27 commonly available instruments used in the evaluation studies were examined. Instruments developed specifically for an individual study were excluded. Data for this descriptive analysis came from primary sources, such as the publisher's web site and the technical manuals accompanying the instruments, and from secondary sources such as the Buros Institute of Mental Measurements (Spies & Plake, 2005). Appendix D provides detailed, publicly available information about the instruments that were used in the evaluations included in the study sample.

The analysis found that the instruments used in the school readiness evaluations reviewed appear

to have been appropriate to the age of children assessed and, for those for which reliability and validity data were available, appear to have generally acceptable reliability and validity data. A number of the instruments did not have the stated purpose of tracking child outcomes or had purposes other than tracking child outcomes. The training recommended to administer the instruments varied, as did the time to administer them.

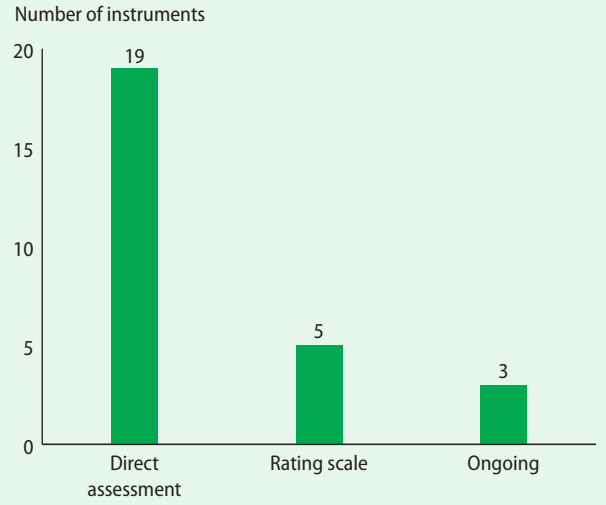
Availability of the instruments. The majority of the instruments used in the evaluations are commercially available through a publishing company. A small group of instruments were developed specifically for research and are not available through a publishing company. Typically, these instruments were developed as part of a large federally funded study such as the FACES Head Start study and then were used in state and local school readiness evaluations. The Color Bears and Counting Bears and Story and Print Concepts assessments are examples. These instruments tend to be tasks presented to children to measure very specific knowledge or skills rather than assessments that measure multiple skills and abilities.

Purpose of the instruments. It is commonly accepted that instruments cannot fulfill multiple purposes and that instruments should be used for the purpose for which they were designed (National Education Goals Panel, 1998). The published data available for the instruments used to collect child outcome data indicated that some instruments were designed solely for that purpose (Color Bears and Counting Bears and Get It! Got It! Go![13]), but the majority were described as having multiple purposes or a purpose other than tracking child outcomes. The instruments with the stated purpose of tracking child outcomes were typically developed for research or program evaluation. Many of the other instruments indicated that they were designed for screening or instruction as well as tracking child outcomes. Five of these (Battelle Developmental Inventory,[14] Comprehensive Test of Phonological Processing,[15] Developmental Observation Checklist System,[16] Expressive One-Word Picture Vocabulary Test,[17] and the Peabody Picture Vocabulary Test–Third Edition) indicated that they were screening instruments, and four (High/Scope Child Observation Record [hereafter referred to as Child Observation Record], Creative Curriculum Developmental Continuum for Ages 3–5 [hereafter referred to as Creative Curriculum],[18] Learning Accomplishment Profile–Revised,[19] and Work Sampling System) indicated that they were for instructional assessments as well as for tracking child outcomes. Four instruments indicated that they were designed to track child outcomes, screen children, and provide instructional assessment information (Oral and Written Language Scales,[20] Preschool and Kindergarten Behavior Rating Scales,[21] Social Skills Rating System, and Woodcock-Johnson III).

A number of the instruments did not have the stated purpose of tracking child outcomes or had purposes other than tracking child outcomes

While these instruments did have a stated purpose of tracking child outcomes, several instruments used in the studies did not. Their primary indicated purpose was screening or instructional assessment. The Get Ready to Read, Preschool Language Scale–Fourth Edition,[22] and Teacher/Child Rating Scale instruments indicated that they are for screening, and the Developmental Indicators for Assessment of Learning–Third Edition[23] indicated that it can be used for screening and for instructional planning but did not list tracking child outcomes as a goal. Four instruments stated they are primarily for instructional assessment—Basic School Skills Inventory–Third Edition,[24] Bracken Basic Concepts Scale–Revised,[25] California Preschool Social Competency Scale,[26] and Developing Skills Checklist. Four instruments—the Preschool Comprehensive Test of Phonological Processing, Pre-Language Assessment Scales,[27] Receptive One-World Picture Vocabulary Test, [28] and Story and Print Concepts—stated their purpose as assessing particular skills, and their purpose could not be classified as screening, instructional, or tracking of child outcomes. While a majority of the instruments included a stated purpose of tracking child

FIGURE 5
Most of the instruments were direct assessments (*N* = 27)



Source: See appendix D.

outcomes, among other purposes, some were not designed to collect data on child outcomes.

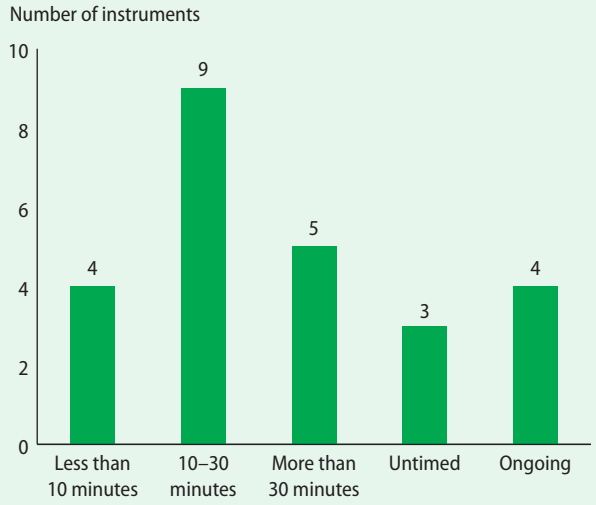
Administration of the instruments. The review also examined data on how the instruments were intended to be administered—how data are collected, the age of children for which the instruments are designed, training required to administer the instrument, and the time needed to administer the instrument.

Most of the instruments (19 of 27) are designed for direct child assessments, with a test administrator presenting specified tasks and recording the child's response (figure 5). Five of the instruments are rating scales, completed by someone who knows the child well and bases the assessment on prior knowledge of the child (Teacher Child Rating Scale, Social Skills Rating System, Preschool and Kindergarten Behavior Rating System, Developing Skills Checklist, and Developing Observation Checklist System). Three of the instruments are "authentic assessments." The assessor collects ongoing data on the child's performance during regular instructional activities and then records the results, typically in the form of a rating scale. The Child Observation Record, Creative Curriculum, and Work Sampling System assessments fall into this category.

The target age range of the instruments varied. Thirteen of the instruments are designed to be used for preschool- and school-age children, primarily kindergartners through second-graders (Basic School Skills Inventory, Bracken Basic Concepts Scale–R, Child Observation Record, Developing Skills Checklist, Developmental Indicators for the Assessment of Learning–3, Expressive One-Word Picture Vocabulary Test, Learning Accomplishment Profile–R, Oral and Written Language Scales, Pre-Language Assessment Scales, Preschool and Kindergarten Behavior Rating Scales, Social Skills Rating System, and the Work Sampling System). Five instruments could be used with a wider age range, from infants and toddlers and through school age or even adulthood (Battelle Developmental Inventory, Developmental Observation Checklist System, Peabody Picture Vocabulary Test, Preschool Language Scale–4, Receptive One-Word Picture Vocabulary Test, and the Woodcock-Johnson III). In contrast, seven instruments were designed to collect data from a narrow age span—preschool-age children only (California Preschool Social Competency Scale, Color Bears and Counting Bears, Creative Curriculum, Get It! Got It! Go!, Getting Ready to Read, Preschool Comprehensive Test of Phonological Processing, and the Stories and Print Concepts). Two instruments (the Comprehensive Test of Phonological Processing and the Teacher/Child Rating Scale) are designed to be used with school-age children, not with preschool-age children.

The average or recommended amount of time to administer the instruments ranged from a few minutes to multiple observations conducted over an extended period (figure 6). Several can be administered in less than 10 minutes (Basic School Skills Inventory, California Preschool Social Competency Scale, Get it! Got it! Go!, and Getting Ready to Read). Several typically require 10–30 minutes (Comprehensive Test of Phonological Processing, Developmental Observation Checklist System, Developmental Indicators for the

FIGURE 6
Time required to administer the instrument varies considerably



Source: See appendix D.

Assessment of Learning–3, Expressive One-Word Picture Vocabulary Test, Peabody Picture Vocabulary Test–III, Preschool and Kindergarten Behavior Rating Scales, Receptive One-Word Picture Vocabulary Test, Social Skills Rating System, and the Teacher/Child Rating Scale). And several typically take more than 30 minutes to administer (Battelle Developmental Inventory, Oral and Written Language Scales, Preschool Comprehensive Test of Phonological Processing, Preschool Language Scale–4, and the Woodcock-Johnson III). Three of the instruments stated that they are untimed and did not provide an estimate of the amount of time required for administration. (Bracken Basic Concepts Scale–R, Developing Skills Checklist, and Learning Accomplishment Profile–R), and three depend on ongoing teacher observations (Child Observation Record, Creative Curriculum, and the Work Sampling System).

A number of the instruments indicated that they are available in Spanish (Battelle Developmental Inventory, Bracken Basic Concepts Scale–R, Child Observation Record, Creative Curriculum, Developing Skills Checklist, Developmental Indicators for the Assessment of Learning–3, Expressive One-Word Picture Vocabulary Test, Getting Ready to Read, Peabody Picture Vocabulary Test–III, Preschool Comprehensive Test of Phonological Processing, Pre-Language Assessment Scales, Preschool and Kindergarten Behavior Rating System, Preschool Language Scale–4, Receptive One-Word Picture Vocabulary Test, Story and Print Concepts, Woodcock-Johnson III, and Work Sampling System). None indicated availability in other languages.

The training recommended for the person administering the instruments varied. Most of the instruments suggested or required training specific to the instrument (Battelle Developmental Inventory, Child Observation Record, Creative Curriculum, Developmental Indicators for the Assessment of Learning–3, Expressive One-Word Picture Vocabulary Test, Getting Ready to Read, Learning Accomplishment Profile–R, Story and Print Concepts, Woodcock-Johnson III, and Work Sampling System). Typically, this training is available through the publisher. Several instruments require college-level training in psychological testing or a related field (Bracken Basic Concepts Scale–R, Oral and Written Language Scales, Peabody Picture Vocabulary Test–III, Receptive One-Word Picture Vocabulary Test, and Social Skills Rating System). Three instruments indicated that someone knowledgeable in child development, child assessment, or early childhood education could administer the instrument (Basic School Skills Inventory, Developmental Observation Checklist System, and Preschool and Kindergarten Behavior Rating System). For the remaining instruments either no level of training was specified or no information was found on recommended qualifications.

Technical information available on the psychometric properties of the instruments. Selecting the best instruments for assessing young children is a difficult challenge for educators. An instrument's track record of providing reliable and valid data is an important consideration. This section summarizes published information on reliability and validity data for each instrument (except the Teacher/Child Rating Scale). Detailed information about each instrument is presented in appendix D.

Instruments for which reliability and validity data were available appear to have generally acceptable reliability and validity data

As already mentioned, it is generally recommended that instruments used in research or to examine child outcomes (but not for decisions about individual children) demonstrate reliability coefficients of .80 or above (Gall, Borg, and Gall, 1996; Cohen & Spenciner, 2007). Most of the instruments for which reliability data were available have examined the internal consistency of the instrument, and all but six of the instruments reported internal consistency reliability coefficients of .80 or greater. In a few cases the reliability coefficients dropped below .80, but this was typically on one subscale rather than the full scale. A few instruments reported internal consistency reliability coefficients of .90 or higher (Basic School Skills Inventory–III, Bracken Basic Concepts Scale–R, California Preschool Social Competency Scale, Color Bears and Counting Bears, Creative Curriculum, Expressive One-Word Picture Vocabulary Test, Peabody Picture Vocabulary Test–III, Preschool and Kindergarten Behavior Rating Scales, and Receptive One-Word Picture Vocabulary Test).

Nine instruments reported data on consistency across raters, or interrater reliability (California Preschool Social Competency Scale, Child Observation Record, Comprehensive Test of Phonological Processing, Developmental Observation Checklist System, Expressive One-Word Picture Vocabulary Test, Learning Accomplishments Profile–R, Oral and Written Language Scales, Preschool and Kindergarten Behavior Rating Scale, and the Preschool Language Scale–4). Interrater reliability coefficients were generally above .80 with the exception of that of the Preschool and Kindergarten Behavior Rating System, which was quite a bit lower.

A third type of reliability examined by instrument developers is test-retest reliability. Fourteen instruments reported this type of reliability (Basic School Skills Inventory, Bracken Basic Concepts Scale–R, Comprehensive Test of Phonological Processing, Developmental Observation Checklist System, Developmental Indicators for the Assessment of Learning–3, Expressive One-Word Picture Vocabulary Test, Get It! Got It! Go!, Learning Accomplishments Profile–R, Preschool and Kindergarten Behavior Rating Scale, Preschool Language Scale–4, Oral and Written Language Scales, Peabody Picture Vocabulary Test–III, Preschool Comprehensive Test of Phonological Processing, and the Social Skills Rating System). The time between administrations ranged from one week to two years. Most test-retest reliability coefficients were .80 or higher. Notable exceptions were the Social Skills Rating System, which reported lower test-retest reliability coefficients for subscales completed by parents (.65) and students (.68) but .80 and higher for those completed by teachers, and the Preschool and Kindergarten Behavior Rating Scale, which reported test-retest reliability coefficients of .58–.86.

The most commonly reported instrument validity is concurrent validity, a comparison of the results of the instrument with results from other instruments that measure similar skills and abilities. Of the 16 instruments that reported concurrent validity studies, results ranged from low to moderate correlations between the results of the instrument being compared. Three instruments (Child Observation Record, Creative Curriculum, and Preschool and Kindergarten Behavior Rating Scales) reported evidence from factor analyses to support the validity of the instruments. Content-validity studies were reported for four instruments (Basic School Skills Inventory, Creative Curriculum, Preschool Language Scale–4, and Social Skills Rating System). Experts reviewed the instrument or completed literature reviews to verify the importance of items contained in the instrument. Information related to the predictive validity of the instruments was found for only three instruments (Bracken Basic Concepts Scale–R, Color Bears and Counting Bears, and Story and Print Concepts).

IMPLICATIONS OF THE FINDINGS RELATED TO INSTRUMENTS USED TO EVALUATE SCHOOL READINESS PROGRAMS

Evaluations of school readiness programs are typically designed to identify the impact of the

programs on child outcomes. Recent reviews of the results of these evaluations have indicated that the program effects on child outcomes are typically small and inconsistent across time periods and domains of children's development (Brown & Scott-Little, 2003; Gilliam & Zigler, 2004).

In discussing similar observations about the research literature that has attempted to document the relationship between the quality of child care settings and child outcomes, Lamb (1998) suggests that limitations in the measurement of child outcomes (and other methodological issues) could contribute to an underestimation of the association between program quality and child outcomes—in short, a lack of precision in how child outcomes are conceptualized and measured could cause studies to inaccurately reflect child outcomes. Therefore, it is important to carefully examine the properties of the instruments and how the data were collected. This section discusses the implications the findings of this project have for evaluations of school readiness programs.

Conceptualization of desired outcomes

Although a criterion for inclusion in this review was that the program being evaluated have the stated goal of improving children's readiness for school success, the complexity of the concepts of school readiness and what constitutes a "successful" outcome result in considerable variability in what and how programs have been evaluated. School readiness is thought of as the degree to which children exhibit certain characteristics or skills associated with later success in school, but there is limited agreement on what characteristics and skills are included in the definition of school readiness, the degree to which they are important, and how to measure them (Snow, 2006).

What is considered important related to school readiness varies from program to program and from person to person (Graue, 2006; Meisels, 1999). Given this lack of agreement, the selection of what outcomes to measure in program evaluations is particularly important. Effective evaluations require that programs have a clear definition of what they are trying to accomplish and subsequently measure. Likewise, decisions about when to measure outcomes have important implications for the conceptualization of school readiness outcomes and the findings that evaluations yield.

A lack of precision in how child outcomes are conceptualized and measured could cause studies to inaccurately reflect child outcomes

Domains of school readiness measured. The evaluations in the studies reviewed have measured outcomes in three broad categories: child development outcomes, school performance outcomes, and long-term outcomes beyond high school. Within the child development outcomes, children's language development and literacy skills were the most frequently targeted outcomes, followed by mathematics/pre-mathematics skills and knowledge. These three areas reflect more academic outcomes. In addition, a substantial number of studies collected data on children's performance on standardized achievement tests once children entered school. Fewer studies assessed social/emotional and health/physical development outcomes.

These findings related to the domains assessed are consistent with Gilliam and Zigler's (2004) review of evaluations of state-funded prekindergarten programs, which found that most program evaluations collected data on a category the authors called "overall development" that included academic skills but was coded separately from "perceived competence," "behavior problems," and "child health" outcomes. In contrast, Zaslow and others (2006) reviewed 65 studies that examined the relationship between children's experiences in child care settings and child outcomes and found that the greatest number of those studies (80 percent) collected outcome data on children's social and emotional development. More than half (54 percent) also collected data on children's cognitive development and general knowledge and on early literacy (51 percent). Based on these data, it seems

that evaluations of school readiness programs in general and prekindergarten programs specifically have more often defined language, early literacy, cognitive development, and general knowledge as important, while studies looking at child care settings have focused more on children's social and emotional development.

Decisions about which program outcomes to examine are an important step in evaluation. There is substantial empirical and theoretical evidence to suggest that children's readiness for school is best conceptualized holistically—that language, early literacy, cognitive development, general knowledge, social and emotional development, and health and physical developmental are all important factors in how likely children are to succeed in school (Huffman, Mehlinger, & Kerivan, 2000; National Education Goals Panel, 1995; National Research Council, 2000; National Research Council, 2001; Snow, 2006). Accordingly, evaluations of school readiness programs should collect data on multiple domains of children's development.

At issue, perhaps, is the availability of valid and reliable instruments across the different domains of school readiness. Overall, measures of cognitive development, language, and literacy often have stronger psychometric properties than measures of other domains of development and learning, particularly social and emotional development and approaches toward learning (Bridges et al., 2004; Denham & Burton, 2003; Snow, 2006; Zaslow & Halle, 2003; Zaslow et al., 2006). However, with the contribution that each area of development makes toward children's overall success in school, there is a need to include appropriate measures in multiple areas of development and to develop or refine measures in nonacademic domain areas.

Many evaluations have collected data on child outcomes at the beginning of kindergarten, but this strategy does not address the need to demonstrate success in school

Time frames for measuring outcomes. In addition to differences in the domains that are included in the conceptualization of school readiness, the program evaluations differed in when they collected data on child outcomes. Decisions about when to collect data must take into account practical considerations, technical considerations about the outcome measures that can be used at various age levels, and conceptual issues related to how the "success" of a school readiness program is defined.

School readiness program evaluators must decide whether to evaluate programs just after children complete the intervention or after children enter school. The largest number of studies elected to follow children into school, but a substantial number collected data only while children were enrolled in the program. This is consistent with Gilliam and Zigler's (2004) observation that most studies of state-funded prekindergarten programs began assessing child outcomes while children were enrolled in the program and continued to follow them once they enrolled in school. The median length of follow-up was second grade, although one study followed students through the tenth grade.

In addition to practical considerations, such as the resources required to conduct longitudinal studies, evaluators should consider the conceptual issues related to how to define program "success" as they make decisions about when to collect data on child outcomes. Collecting data while children are enrolled in the program or at the end of the intervention can address the immediate outcomes of the program. If a strong research design such as an experimental study is used, the evaluation can answer the question "How did the program affect children who were enrolled compared with those who were not?" Although this is an important question, if participants have not yet entered school it is difficult to make statements about their readiness for school.

Many evaluations have collected data on child outcomes at the beginning of kindergarten, when it is easier to locate a comparison group. This strategy does not address the need to demonstrate success in school since children are just beginning kindergarten, and the strategy has the

added complication that there is often a gap in services between the time that children complete the school readiness program and the time they begin kindergarten, a time during the summer when they are receiving no services. During this time any number of variables might affect what children take with them from the school readiness program. Research on Georgia's prekindergarten program has documented an uneven "summer learning loss," with economically disadvantaged African American children more likely to exhibit reduced developmental gains during the summer (Henry, Gordon, Henderson, & Ponder, 2003; Ponder, Rickman, & Henry, 2004).

Another option for collecting data is to follow the children into their early or middle school years. This option provides data related to one of the most significant questions for school readiness program evaluations—did the program help children to be more successful once they entered school? This option too is marred by complex problems that must be addressed to provide a valid assessment of the effectiveness of school readiness programs. First, many instruments that are used with preschool children are not designed to be used with older children. Longitudinal studies therefore often collect data using one set of instruments when children are enrolled in the program and a different set when the children are older. The comparability of results across time must be given careful consideration. In addition, if the program is effective in increasing children's success in school, a number of participants who would otherwise have been referred for special education services may perform well enough to be mainstreamed into the regular student population. Although this is a successful outcome for the individual students, it may result in collectively lower scores over time in the group of children who participated in the school readiness intervention.

Instruments used in the evaluations

Findings related to the types of instruments used, the stated purposes of the instruments, the appropriateness of the instruments for children from a variety of cultural and language groups, and the psychometric properties of the instruments also have important implications for evaluation.

How data are collected. Several types of instruments were used in the evaluations to assess child outcomes, including observational measures, standardized measures, direct assessments, checklists, teacher ratings, and teacher surveys. Each approach has advantages and disadvantages.

Outside observers may provide a more objective, reliable evaluation of children's performance, while teachers and parents have more information about the child and are familiar to the child, but their abilities to collect data vary

Standardized measures and direct assessments often have higher reliability and yield results that are more comparable across children. Yet standardized measures and direct assessments may be less authentic representations of children's abilities. Observational measures often collect more authentic data related to children's performance but may be more time consuming and less reliable, since they are typically conducted by the child's teacher and teachers have varying levels of training and skills needed to conduct assessments. Checklists and rating scales are less expensive and yield results that are comparable across groups of children, but results may reflect any number of factors that can bias the person completing the inventory, ranging from the length of time the person has known the child to the person's own perceptions of what is "typical" development and learning at this age to the settings in which the person has had the opportunity to observe the child.

Similarly, the interaction between the child and the person who completes the instrument must be considered. The majority of the studies used data collected by teachers, but a sizable number used outside observers. Outside observers may provide a more objective, reliable evaluation of children's performance, but children may respond in atypical ways when they are assessed by an unfamiliar

adult. Teachers and parents have more information about the child and are familiar to the child, but their abilities to collect data on child outcome measures vary. They also typically have different levels of information about a child in different situations and settings (Meisels, Bickel, Nicholson, Zue, & Atkins-Burnett, 2001; Schweinhart, 2003).

Sources of instruments used. Most of the evaluations used widely accessible instruments, although a few used study-specific instruments. While study-specific instruments may be more closely tied to program objectives and emphases and therefore may be a more valid measure of program outcomes, the lack of data on reliability and validity is a concern. The use of outcome measures that have not been validated severely limits confidence in the results (Gilliam & Zigler, 2004). For this reason program evaluators should consider using instruments with proven reliability and validity. If they elect to use nonvalidated instruments, they should collect and report reliability and validity data on the instrument as part of the development process.

Evaluators are urged to apply considerable caution when using instruments designed for other purposes to collect child outcome data and to avoid doing so when possible

Among the most commonly used of the widely available instruments were the Peabody Picture Vocabulary Test–III, the Woodcock-Johnson III, the Child Observation Record, and the Social Skills Rating Scale. Several of these instruments were reported to have been used frequently in evaluations both of prekindergarten programs (Gilliam & Zigler, 2004) and of child care quality (Zaslow, et al., 2006). There appears to be some convergence toward a "short list" of child outcome measures, although considerable variability remains.

A key finding in this review of state- and local-level evaluations is the overlap between the instruments used in these evaluations and those used in large federally funded studies. Federal studies of note include the National Institute of Child Health and Human Development Study of Early Child Care and Youth Development, Head Start research (including the Family and Child Experiences Survey, the Head Start Impact Study, and research on the Early Head Start program), and the National Center for Education Statistics Early Childhood Longitudinal Study (see box 4 at the end of this report). In addition to the results of the federal research projects, these projects have made technical data and in many cases the assessment instruments developed by the study available for use by other researchers. Instruments such as the Color Bears and Counting Bears assessment, which have been developed or modified for use in federal studies, or the Peabody Picture Vocabulary Test–III and the Woodcock-Johnson III assessment, which are commercially available instruments, have perhaps been used more frequently in state and local evaluations because of their widespread use in federal studies.

The federal government has invested significant resources in the development and selection of instruments to measure child outcomes, and evaluators for state and local school readiness programs appear to have followed their lead. This may be because of the resources already invested in developing them, the data available on their technical properties, or the possibility of comparing results with findings from a national sample (Kisker et al., 2003).

Purposes of instruments. Many of the instruments that were used to track child outcomes were originally designed for clinical purposes. One reason for the relatively frequent use of instruments designed for other purposes is likely the limited number of instruments designed specifically to track child outcomes. Evaluators are urged to apply considerable caution when using instruments designed for other purposes to collect child outcome data and to avoid doing so when possible. As Snow (2006, pp. 21–22) points out, "Although we may desire a measure that can provide reliable and valid data . . . to serve various purposes of assessment . . ., there are very few measures that can do so, while not sacrificing reliability or validity in the process. It remains to be seen in

early childhood if single measures can inform individual child placement, instruction and serve accountability purposes."

Appropriateness of instruments for children from different cultural and language backgrounds. The number of children from homes where English is not the primary language is growing. Program evaluations should use assessments that are culturally and linguistically appropriate to the children being assessed (Grinder & Kochanoff, 2006; Li, Walton, & Nuttall, 1999). This means that evaluators should seek comparable instruments that are available in the languages that the children speak and have been validated as culturally relevant for the particular group of children. Within the sample examined for this study, 17 instruments are available in Spanish. None of the instruments is available in languages other than Spanish. A very limited number of the evaluation studies, however, reported having used the Spanish version of an instrument.

Unfortunately, the number of culturally and linguistically appropriate instruments that can be used with non-English-speaking children to provide data that is comparable to instruments administered in English is limited. Even when a Spanish version of an instrument is available, it may not be equivalent to the English instrument. For example, the Spanish vocabulary test Test de Vocabulario en Imagenes Peabody, produced by the same publisher as the Peabody Picture Vocabulary Test–III, which was used in many of the studies reviewed, is not the same test but is based on an earlier version of the English test (Pearson Assessment, 2007). Therefore, data collected using the two assessments are not equivalent and should be analyzed separately.

Evaluators must also establish that the norming group for non-English assessments is appropriate to the children who will be assessed (Demers & Fiorello, 1999; Kisker et al., 2003). For instance, many instruments that are available in Spanish have been normed with native Spanish speakers in other countries, such as Mexico or Spain. The norms for these instruments would not be appropriate for Spanish-speaking children growing up in the United States. Furthermore, even if a Spanish-language assessment was normed with Spanish-speaking children in the United States, evaluators must understand the particular population that was used for the norming group and the extent to which the group is representative of the children they will be assessing.

Finally, evaluators should include a valid and reliable screening instrument in the research protocol to determine whether English language learners are sufficiently proficient in English to be assessed in English and, if assessments are administered in English, whether there are items that might be culturally biased or unfamiliar to the group of children being assessed. Li and others (1999) provide a review of various types of instruments that have been used with culturally diverse preschoolers.

Psychometric properties of the instruments. Ensuring that instruments used to evaluate school readiness programs are valid and reliable for the specific children being assessed is particularly important because of the high degree of variability in the learning and development of children at this age and the limited number of appropriately normed and validated instruments for very young children (Demers & Fiorello, 1999).

Reliability data were located for 26 of the 27 instruments most commonly used in the evaluations. Evidence from validity studies was located for 23 instruments. For the instruments for which data were available, reliability and validity were largely acceptable. Most instruments reported tests of internal consistency, and several reported other types of reliability. For several instruments studies were conducted comparing results with those of other assessments completed on the same children to examine the

Evaluators should consider the types of reliability data available, the ages of the children from whom the reliability data were collected, and the results of reliability studies

validity of the instrument. A number of instruments, particularly instruments designed specifically for individual studies, lacked reliability or validity data or both. These findings echo those of Zaslow and others (2006, pp. 594–95), who state, "Whereas we do not see serious and pervasive problems with inadequate reliability or validity [in the measures where data were located], the lack of reporting of psychometric information on so many measures, particularly in the domain of socio-emotional development, may mask underlying problems with the strength of the measures." Evaluators should consider the types of reliability data available, the ages of the children from whom the reliability data were collected, and the results of reliability studies to ensure that an instrument demonstrates adequate reliability with the sample that will be assessed in an evaluation.

One issue worthy of note is the lack of evidence of predictive validity. Evidence of predictive validity was located for only 3 of the 27 commonly used instruments. Results from several previous meta-analyses have suggested that instruments used with very young children have limited predictive validity and therefore that it may not be appropriate to draw long-term conclusions about children's development and learning from assessments given to very young children (Kim & Suen, 2003; La Paro & Pianta, 2000). Overall, measures of cognitive or academic skills seem to have a higher degree of predictive validity than measures of children's social and emotional development, and tests scored through ratings were more predictive than other types of assessments. This could be due to a variety of factors, including the instability of children's development and learning at this age and properties of the measures themselves.

Predictive validity is particularly important for instruments being used in evaluations of school readiness programs

Predictive validity is particularly important for instruments being used in evaluations of school readiness programs. It is important that the instruments assess children's skills and knowledge that are associated with positive child outcomes later in life. It is difficult for test developers to collect data on predictive validity because this requires longitudinal research. Data on the predictive validity of widely used instruments may be available in the larger research literature. To select instruments that have a proven track record—that have established reliability and validity and are likely to have predictive validity—evaluators may find it helpful to conduct a literature search of studies that have used a particular instrument and to contact the test developer to request information on how the instrument has been used in longitudinal research and how it has demonstrated predictive validity.

Administration of the instruments

Assessments that are valid, reliable, and culturally and linguistically appropriate can still produce invalid results if they are administered incorrectly. It is therefore important to pay attention to who is administering the assessment and how well prepared that person is to administer the assessment.

Data from this review suggest that teachers are by far the most likely people to administer the assessments. They did so in 45 of the 65 assessments used in these evaluations. Having teachers collect program-evaluation data may have both practical and theoretical justification. From a practical standpoint it is often more economical to have teachers complete assessments—they are with the children regularly and can complete the assessments at minimal additional cost. From a theoretical perspective, teachers can collect data in a more naturalistic fashion, they have access to a variety of types of data about children, and they are familiar to the children so they may be more likely to see a wider sample of children's skills and abilities (Snow, 2006).

Some question the use of teachers as data collectors, however. Where teacher judgment of student learning is a key element of performance assessment, some researchers have expressed concerns about the validity and reliability of teacher

assessment (Hoge & Coladario, 1989; Sattler, 2002), particularly in high-stakes situations such as program evaluations. However, others (Meisels and others 2001; Schweinhart, 2003) maintain that teachers can collect data that are both reliable and valid.

Little information is available about what kind of preparation teachers and other data collectors received to collect assessment data in a reliable and accurate manner for these program evaluations. Only about half of the studies even mentioned that the data collectors were trained, and very few described the training. Only four studies documented the training and demonstrated reliability in administering the scale. Fifteen of the 17 commonly used instruments specified that people who administer them should have formal training in the assessment instrument, and 5 recommended that assessors also have college-level training in assessment or psychometrics. These issues highlight the need to ensure that data collectors are adequately trained to collect data in a consistent and standardized way that permits comparisons across classrooms or programs.

When considering instruments, evaluators should carefully study the test's administrative procedures and the training recommended by the authors or publishers to ensure that the evaluation will be able to follow the recommended procedures. Instrument authors often produce procedures manuals that describe how an instrument should be administered. For instruments developed for and used in large federally funded studies, procedures manuals are often publicly available or can be accessed by contacting the instrument author (see box 4). Commercially available instruments often have procedures manuals and specialized training available for purchase.

Evaluators should consider how an instrument is designed to be administered to determine whether it is appropriate for a particular evaluation and the people who will be collecting the data. Once an instrument is selected, training should be provided for the data collectors to ensure that procedures are carried out appropriately. Evaluators could consider a process for certifying that the data collectors are properly trained, which could include specified training and a requirement that data collectors demonstrate their ability to administer the instrument in accordance with the test's specified procedures prior to collecting data. Once data collection has begun, data should be collected to establish that data collectors are assessing children reliably and accurately.

Once an instrument is selected, training should be provided for the data collectors to ensure that procedures are carried out appropriately

A final consideration related to instruments is the burden that the data collection process will place on children and staff in the program being evaluated. Assessing children and conducting evaluations can be challenging and time consuming. For example, in Georgia's early childhood study (Henry et al., 2004, 2005) a battery of 13 different assessments was administered to approximately 800 children during the study (fall 2001–spring 2004). Typically, children were administered 6–8 assessments at each point in the data collection, with data being collected by both teachers and outside assessors. Putting together such an assessment is a formidable and expensive task. Carrying it out is also a burden on teachers and children, one that needs to be considered in the context of other assessment requirements that teachers and children face as a result of other program mandates (such as the Head Start National Reporting System [Administration for Children, Youth, & Families, 2003] or the Early Childhood Outcomes requirements for Individuals with Disabilities Education Act [Hebbler, Barton, & Malik, 2007]) or in the course of program instruction (screening and instructional assessments). Without careful planning and coordination teachers and children can spend an inordinate amount of time completing assessments. As decisions are made about how to collect data, evaluators must consider the resource costs and how much time will be taken from instruction.

The need to collect data related to the child's environment

Recent definitions of school readiness have emphasized the fact that factors such as the learning environment, the child's family, and characteristics of the school where the child is enrolled are important components of "readiness" (Graue, 2006; Snow, 2006). Thus while evaluators might be interested primarily in child outcomes, a program evaluation must collect data on other factors in order to know to what extent the child outcomes are likely to be effects of the program itself.

Thirty-one of the studies reviewed collected both implementation and summative data, while 45 collected only summative information on child outcomes. That is, less than half of the studies collected any information on the nature of the program in which the children were enrolled. It is important that researchers look beyond child outcomes to program quality and classroom dynamics. One cannot fully understand child effects without first considering the quality of the program. Absent a measure of quality, if a program fails to demonstrate positive results, it is difficult to determine whether the results are due to poor design or poor implementation (Gilliam & Zigler, 2001).

Similarly, it is important for evaluations to collect data on children's family backgrounds and, if they follow children longitudinally, on the schools in which they enroll. Only with a complete picture of children's family backgrounds, the school readiness intervention they received, and the schools in which they are enrolled can evaluators have a true picture of a program's impact.

BOX 4

Resources available for making decisions about child assessments

Resources to guide decisionmaking

Grinder, E. L. & Kochanoff, A. (2006). *Updated report on early childhood assessment for children from birth to age 8 (grade 3).* Harrisburg, PA: Pennsylvania's Departments of Education and Public Welfare. Retrieved from http://www.pde.state.pa.us/early_childhood/lib/early_childhood/Build_9-06_AssessB-8.pdf

Kisker, E. E., Boller, K., Nagatoshi, C., Sciarrino, C., Jethwani, V., Zavitsky, T., Ford, M., & Love, J. M. *Resources for measuring services and outcomes in Head Start programs serving infants and toddlers*. Princeton, NJ: Mathematica Policy Research, Inc., April 2003. Retrieved from http://www.acf.hhs.gov/programs/opre/ehs/perf_measures/reports/resources_measuring/res_meas_title.html

National Education Goals Panel. (1998). *Principles and recommendations for early childhood assessments*. Washington, DC: Author. Retrieved from http://govinfo.library.unt.edu/negp/Reports/prinrec.pdf

Nuttall, E. V., Romero, I., & Kalesnik, J. (1999). *Assessing and screening preschoolers: Psychological and educational dimensions (2nd edition).* Needham Heights, MA: Allyn & Bacon.

Ross, C., Kirby, G., Schochet, P., Hall, J., Sprachman, S., Boller, K., Paulsell, D., & McConnell, S. (2005). *Design options for the assessment of Head Start quality enhancements. Final report, Vol. 1.* Princeton, NJ: Mathematica Policy Research, Inc.

Scott-Little, C., Kagan, S. L., & Clifford, R. M. (Eds.). (2003). *Assessing the state of state assessments: Perspectives on assessing young children*. Tallahassee, FL: SERVE.

Scott-Little, C., & Niemeyer, J. A. (2001). *Assessing kindergarten children: What school systems need to know.* Tallahassee, FL: SERVE.

Reviews of specific instruments

Bridges, L. J., Berry, D. J., Johnson, R., Calkins, J., Margie, N. G., Cochran, S. W., Ling, T. J., & Zaslow, M. J. (2004). *Early childhood measures profiles.* Washington, DC: Child Trends. Retrieved from http://aspe.hhs.gov/HSP/ECMeasures04/index.htm

Kisker, E. E., Boller, K., Nagatoshi, C., Sciarrino, C., Jethwani, V., Zavitsky, T., Ford, M., & Love, J. M. (2003, April). *Resources for measuring services and outcomes in Head Start programs*

CONCLUSIONS AND RESOURCES

Evaluations of school readiness programs are increasingly common. The following recommendations based on the data collected from this sample of school readiness evaluations are provided to guide school readiness programs and evaluators as they select and implement child assessments:

- Carefully select outcomes for assessment that match the goals of the program and that address the components of children's learning and development that are linked with later success in school.
- Clearly define the purpose for which the assessment data will be collected, and select instruments that have been designed and validated for that purpose.
- Select instruments that have a proven track record with children who have the characteristics of those who will be assessed (instruments that have adequate reliability and validity and that have been tested with children similar to those served by the program).
- Select instruments that are culturally and linguistically appropriate for the children who will be assessed.
- Consider whether outside observers or people who work directly with the children are the best collectors of data.

serving infants and toddlers. Princeton, NJ: Mathematica Policy Research, Inc. Retrieved from http://www.acf.hhs.gov/programs/opre/ehs/perf_measures/reports/resources_measuring/res_meas_title.html

Li, C., Walton, J. R., & Nuttall, E. V. (1999). Preschool evaluation of culturally and linguistically diverse children. In E. V. Nuttall, I. Romero, & J. Kalesnik (Eds.). *Assessing and screening preschoolers: Psychological and educational dimensions (2nd ed.).* Needham Heights, MA: Allyn & Bacon.

Pennsylvania Department of Education. (2006). Assessment tools. Retrieved from http://www.pde.state.pa.us/early_childhood/cwp/view.asp?A=179&Q=101706

Pai-Samant, S., DeWolfe, J., Caverly, S., Boller, K., McGroder, S., Zettler, J., Mills, J., Ross, C., Clark, C., Quinones, M., & Gulin, J. (2005). *Measurement options for the assessment of Head Start quality enhancements.* (Vol. II). Princeton, NJ: Mathematica Policy Research, Inc.

Information on measures used in select federally funded research studies

Chapman, C., Germino-Hausken, E., Mulligan, G. M., Park, J., & Tice, P. (2006). Early Childhood Longitudinal Study (funded by the Institute of Education Sciences, National Center for Education Statistics). Retrieved from http://nces.ed.gov/ecls/Birth.asp and http://nces.ed.gov/ecls/Kindergarten.asp

Early Head Start Research and Evaluation Project (funded by the Administration for Children and Families, Department of Health and Human Services). Retrieved from http://www.acf.hhs.gov/programs/opre/ehs/ehs_resrch/index.html

Head Start Family and Children's Experiences Survey (funded by the Administration for Children and Families, Department of Health and Human Services). Retrieved from http://www.acf.hhs.gov/programs/opre/hs/faces/overview

Head Start Impact Study (funded by the Administration for Children and Families, Department of Health and Human Services). Retrieved from http://www.acf.hhs.gov/programs/opre/hs/impact_study/index.html

National Institute of Child Health and Human Development Study of Early Child Care and Youth Development (funded by the National Institute of Child Health and Human Development). Retrieved from http://www.nichd.nih.gov/research/supported/seccyd.cfm

- Plan carefully for how the assessments will be administered, provide adequate training for data collectors, and carry out reliability studies to determine whether the data are being collected reliably and accurately.

- Collect data on the children's home context, the nature of the school readiness program in which the children are enrolled, and (if collecting data once children enter school) on the school in which the children are enrolled.

Selecting and implementing instruments for evaluating school readiness programs are no easy task. The findings of this report highlight the challenges that evaluators face in ensuring that data are collected in a manner that yields credible, trustworthy, and meaningful information about child outcomes. There are, however, a number of useful resources available to guide evaluators in making decisions about child assessments. Box 4 lists three types of resources: resources to guide decisions about how to assess child outcomes, reviews of instruments, and web sites with technical information related to instruments used in large federal studies. With careful study of these and other resources, thoughtful planning, and vigilant implementation, evaluations can yield credible data to gauge the outcomes of school readiness programs.

NOTES

The authors would like to extend a special thanks to the following individuals who provided information, feedback, and support in the development of this report: Donna Bryant and Frank Porter Graham, Child Development Institute, University of North Carolina, Chapel Hill; Amy Detgen, Lynn Gregory, Dana Holland, and Richard Sawyer, Academy for Educational Development; and Art Hood, Wendy McColskey, and Kathleen Mooney, SERVE Center, University of North Carolina, Greensboro.

1. Whitehurst, G. (2006).
2. High/Scope Educational Research Foundation (1992).
3. Hightower, D., Work, W., Cowen, E., Lotyczewski, B., Spinell, A., Guare, J., & Rohrbeck, C. (1986).
4. Dunn, L. M., & Dunn, L. M. (1997).
5. McGrew, K. S., & Woodcock, R. W. (2001).
6. The Woodcock-Johnson III consists of 22 subtests. In most instances the evaluations reviewed for this study used only selected subtests, not the full battery of subtests. The most commonly used subtests were Applied Problems, Letter-Word Identification, Math Fluency, and Sound Awareness.
7. Gresham, F. M., & Elliott, S. N. (1990).
8. Kisker, E. E., Boller, K., Nagatoshi, C., Sciarrino, C., Jethwani, V., Zavitsky, T., Ford, M., & Love, J. M. (2003).
9. Kisker et al. (2003).
10. Meisels, S., Jablon, J., Dichtelmiller, M., Dorfman, A., & Marsden, D. (1998).
11. Lonigan, C. J. (2006).
12. CTB (1990).
13. Early Childhood Research Institute on Measuring Growth and Development (1998).
14. Newborg, J., Stock J., Wnek, L., Guidubaldi, J., & Svinicki, J. (1984).
15. Wagner, R., Torgeson, J., & Rashotte, C. (1999).
16. Kisker et al. (2003).
17. Bridges et al. (2004).
18. Dodge, D., Colker, L., & Heroman, C. (2000).
19. Chapel Hill Training and Outreach Project (1995).
20. Carrow-Woolfolk, E. (1996).
21. Merrell, K. W. (1996).

22. Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2002).

23. Cizek, G., & Fairbanks, D. (2004).

24. Hammill, D., Leigh, J., Pearson, N., & Maddox, T. (1998).

25. Bracken, B. A. (1998).

26. Levine, S., Elzey, F. F., & Lewis, M. (1969).

27. Duncan, S.E., & De Avila, E. (1998).

28. Brownell, R. (2000).

APPENDIX A
METHODOLOGY

Identifying and selecting relevant studies was a multistage process, beginning with a literature search and ending with a sample of 82 evaluation documents covering 78 studies.

Conducting the search

To identify relevant studies and reports, a list of key words was compiled that might be used to describe school readiness initiatives. Search terms included *early intervention, school readiness, Head Start, literacy, prekindergarten,* and *kindergarten readiness*. These terms were combined with key words such as *outcomes, evaluation*, and *results* to search specifically for studies that relate to program evaluations and child outcomes.

First, an extensive search was conducted of state department of education web sites to locate any information and evaluations related to a state's school readiness initiatives. The World Wide Web was also searched using the terms outlined above. This led to such web sites as Child Trends, Mathematica Policy Research, National Center for Early Development and Learning, the RAND Corporation, the regional educational laboratories of the U.S. Department of Education, Southern Regional Education Board, and U.S. Department of Health and Human Services. Links from these sites were followed and searched thoroughly. Finally, conference programs and proceedings were reviewed for the American Educational Research Association, Head Start National Research Conference, National Association of Elementary School Principals, and the Society for Research in Child Development.

In addition to the literature review, early childhood specialists in each state department of education were asked about any evaluations of their state's prekindergarten program. Specialists were contacted through meetings, email, and email lists and asked for any evaluation reports that might document the effectiveness of their programs.

The search was conducted between May and July 2006 and yielded 217 articles, reports, conference presentations, and dissertations that had been disseminated between January 1997 and July 2006. Titles and abstracts identified through the search were screened for relevance using the inclusion criteria identified below. The remaining documents were carefully reviewed to determine which studies might be appropriate to include in the review. Documents that did not report on child outcomes were eliminated, as were reports of systems designed to provide a "snapshot" of children as they entered school (such as programs in North Carolina, Maryland, and Vermont). In addition, multistate reports and evaluations of federal programs such as Head Start and Even Start were eliminated. Based on this initial screening, a total of 148 documents were identified as candidates for inclusion.

Narrowing the pool

Selection criteria were developed related to the nature of the program being evaluated, along with some basic requirements for the evaluation report, to decide which reports to include. Documents meeting the following criteria were included in the next stage of review:

- *Recent publication date.* Reports or studies published between January 1997 and July 2006 were included, since a goal of this review was to examine data from recently reported evaluations.

- *Programs located in the United States.* Only evaluations of school readiness initiatives located in the United States were included.

- *Publicly funded programs or interventions that directly target children from birth to age five and whose stated goal is to enhance children's readiness for school.* In addition, programs must maintain a component of classroom-based services. Programs that used solely a home-visitation or parent-education approach were not included. Neither were drop-in or child care programs.

- *Evidence of effectiveness.* The report must present some type of child outcome data (such as achievement test scores, retention rates, behavior problem referrals, special-education placement) indicating the effectiveness of the program.

- *Sufficiency of information.* The report must provide sufficient information for the study to be coded.

Once all of the potentially relevant studies had been identified, the full version of each study was retrieved and reassessed for whether it met the inclusion criteria. Of the 148 documents evaluated, 82 met the inclusion criteria.

Coding of the studies

Each of the remaining documents was coded using a detailed system for capturing a general description of the school readiness program and the methods used to evaluate its effectiveness (see appendix B). The coding sheet included information about the location and setting of the program and the type of intervention (such as state-funded prekindergarten or general school readiness initiative). For evaluation of the program, data were collected on the type of intervention (implementation, summative, or both), the time frame and procedure for data collection, the outcomes assessed, and the instruments used to collect the outcome data.

People experienced in research and evaluation served as coders. Each coder received training and coded a sample of studies. Based on this process, the coding sheet was refined and coding instructions were clarified. Each coder was assigned specific studies to code. Reliability was established by having the project coordinator verify the codings for a sample of studies (10 percent). The data from the coding sheets of the data coder and the project coordinator were compared for agreement. The raters attained 100 percent agreement on key study variables including type of evaluation conducted, data collection timeframe, and number of instruments.

Coding commonly used instruments

Based on the results of the coding process, each instrument used in the evaluation studies was identified. A separate coding process was then used to collect information on each instrument. This coding process was developed to capture a general description of the assessment, the purposes of the assessment, and the target population. In addition, information was summarized about how each instrument is administered, along with information on the scores and scoring procedures. This coding process also collected publicly available technical information on validity and reliability for each instrument. This information was taken from materials provided with the assessment instrument, and, in most instances, from the publisher's or distributor's web site. Information was also collected from other sources such as the Buros Institute of Mental Measurements (Spies & Plake, 2003) and the Administration for Children, Youth, and Families of the U.S. Department of Health and Human Services (Kisker et al., 2003). A summary of the information collected about the more commonly used assessment instruments is in appendix D.

Completing the sample

The final sample for this review included 82 evaluation documents covering 78 separate studies. A study was defined by the population sample. For example, reports that continued to follow a specific cohort of children over time (such as the Georgia Prekindergarten Program and Washington's Early Childhood Education Assistance Program) reflect a single study, whereas a program that used the same methodology and dependent variables but added new cohorts each year was coded as multiple studies.

The sample included reports published between 1997 and the first half of 2006. Before 2002 only a handful of programs had published evaluation reports designed to document program improvement and accountability. Since 2002 the number of evaluation reports that have been made publicly available has increased dramatically, perhaps

reflecting the increasing emphasis on documenting program outcomes or the movement to hold programs accountable.

The 82 reports reviewed 41 separate programs covering 26 state-funded readiness initiatives. Most are administered by a state department of education or another state agency such as Georgia's Department of Early Care and Learning or North Carolina's Partnership for Children. The remainder are local programs supported by other funding, such as federal Title I funding.

Within the 41 programs children were served in school-based and community-based settings. At least 11 programs maintain classrooms primarily in school settings, while 6 programs serve children primarily in community-based sites. Nineteen programs serve children in classrooms established in both school-based and community-based sites. The remaining five programs were coded as "unspecified," either because the information was not presented in the report or because the information was insufficient to determine the setting.

APPENDIX B
CODING SHEET AND EVALUATIONS REVIEWED

School readiness evaluation project study coding sheet

Items marked with an asterisk are items for which we would establish reliability.

REPORT IDENTIFICATION

Report Code No.: _____

Study Code No.: _____

Report Year/Date: _____

Program Name: _____

Title of Report: _____

CHARACTERISTICS OF THE PROGRAM/INTERVENTION

Location/Implementation of the Program

Name of City(ies)/District(s): _____
_____

Name of State: _____

Type of Program
- ☐ 1. State
- ☐ 2. Local
- ☐ 3. Unknown

PROGRAM OR INTERVENTION INFORMATION

Intervention Type
- ☐ 1. State-funded pre-kindergarten
- ☐ 2. Head Start
- ☐ 3. Even Start/Family Literacy
- ☐ 4. Early Reading First
- ☐ 5. Child Care
- ☐ 6. Other: _____
  _____

Type of Setting (location of program)
- ☐ 1. School
- ☐ 2. Community-based
- ☐ 3. Mixed
- ☐ 4. Not specified

EVALUATION STUDIES INFORMATION

Type of Evaluation*
- ☐ 1. Implementation (i.e., formative or process)
- ☐ 2. Summative
- ☐ 3. Implementation and summative

Data Collection Timeframe
- ☐ One time while subjects in program
- ☐ Multiple times while subjects in program
- ☐ Only after subjects complete program
- ☐ While subjects in program and after they leave the program

RESEARCH DESIGN FOR OUTCOME DATA

1. Design Type

- ☐ A. Pseudo-experimental
  - ☐ 1. One group or case study
  - ☐ 2. One-group pretest–post-test design
  - ☐ 3. Static group comparison design
  - ☐ 4. Other: _____
    _____

- ☐ B. Quasi-experimental
  - ☐ 1. Nonequivalent control group, post-test only

    Matching
    - ☐ 1. Yes
    - ☐ 2. No

    Matched on: _____

    Statistical controls used for analysis
    - ☐ 1. Yes
    - ☐ 2. No

    Variables used as controls: _____

  - ☐ 2. Nonequivalent control group, pre-test/post-test

Matching
- ☐ 1. Yes
- ☐ 2. No

Matched on: _____

Statistical controls used for analysis
- ☐ 1. Yes
- ☐ 2. No

Variables used as controls:
- ☐ 3. Correlational
- ☐ 4. Time series
- ☐ 5. Other: _____

_____

☐ C. Experimental (with random assignment)
- ☐ 1. Post-test only control group
- ☐ 2. Pre-test–post-test control group
- ☐ 3. Solomon Four Group
- ☐ 4. Other

☐ D. Not applicable (process/formative evaluation only)

MEASUREMENT OF DEPENDENT VARIABLES

Outcomes Measured within This Study:
- ☐ Cognitive development
- ☐ School achievement (e.g., grades)
- ☐ General knowledge and awareness
- ☐ Delinquency/arrests
- ☐ Math/pre-math
- ☐ Retention
- ☐ Social-emotional development
- ☐ Special-education services
- ☐ Physical development
- ☐ School attendance
- ☐ Language/communication
- ☐ Dropout of school
- ☐ Literacy/pre-literacy
- ☐ Expulsion
- ☐ Overall development
- ☐ Dropout of school
- ☐ Self-help/functional performance
- ☐ Standardized test scores
- ☐ School adjustment/attitude
- ☐ Other: _____

Number of Instruments Used? _____

Were the following specific measures used? (check all that were used)
- ☐ Ages and Stages Questionnaire
- ☐ Battelle Developmental Inventory
- ☐ Bracken Basic Concept Scale–Revised
- ☐ Brigance
- ☐ California Preschool Social Competency Scale
- ☐ Carolina Developmental Inventory
- ☐ Child Observation Record
- ☐ Color Bears
- ☐ Color Names and Counting
- ☐ Counting Bears
- ☐ Creative Curriculum
- ☐ Comprehensive Test of Phonological Processing
- ☐ Developmental Observation Checklist System
- ☐ Developmental Indicators for the Assessment of Learning–Revised
- ☐ Developmental Indicators for the Assessment of Learning–Third Edition
- ☐ Indicators of Basic Early Literary Skills
- ☐ Get It! Got It! Go!
- ☐ Get Ready to Read
- ☐ Learning Accomplishment Profile–Diagnostic Edition
- ☐ Letter Identification and Concepts About Print
- ☐ Oral Written and Language Scales
- ☐ Peabody Picture Vocabulary Test–Third Edition
- ☐ Pre-Language Assessment Scales 2000
- ☐ Social Skills Rating System
- ☐ Story and Print Concepts
- ☐ Woodcock-Johnson III
- ☐ Work Sampling System (or a variation)
- ☐ Measure developed specifically for this study:

_____

List all other measures: _____

_____

_____

_____

_____

Additional notes/comments: _____

_____

_____

_____

Summary of child outcome measures and how they were used										
Name of measure	Construct/ outcome(s) it measures	Reported use(s) of measure[a]	Adaptations made? (Y or N— If Y, please describe below)[b]	Subscale or complete measure (S or C)— list which subscales were used	Non-English version used? (Y/N)[b]	Number of subjects administered assessment?	Who provided the data? (teacher, outside observer, others)	Reported training for data collector (Y/N—if Y, describe below)[b]	Age/ grade level measure used with[c]	Reliability/ validity data from this study

a. Purposes: screening (S), instructional (I), tracking child outcomes (CO), research purposes (R), purpose not specified (NS), other (O).

b. Y = Yes (list language and version below), N = No.

c. Infant/toddler (birth to two); three-year-old preschoolers; four-year-old preschoolers; five-year-old preschoolers; kindergartner; first grade; second grade; third grade; fourth grade and above.

Adaptations made for measures (describe by measure):

_____
_____
_____
_____
_____

Non-English version(s) used: _____

_____
_____
_____
_____

Subscales used (list by measure): _____

_____
_____
_____
_____

Training provided for data collectors (list measure and describe): _____

_____
_____
_____
_____

Results (include effect sizes if available)

Outcome measure	Results	Comments (include whether results were statistically significant)
a.		
b.		
c.		
d.		
e.		
f.		

Coder Name: _____

Date: _____

APPENDIX C
EVALUATION REPORTS REVIEWED, BY STATE AND PROGRAM

Arizona

At-Risk Preschool Program

Norton, D. (1997). *An evaluation of the At-Risk Preschool Program*. Phoenix, AZ: Office of the Auditor General.

California

First 5 California

California Children and Families Commission. (2003). *First 5 school readiness initiative pilot study: Overview and results.* Sacramento, CA.

California Children and Families Commission. (2004). *First 5 school readiness initiative: Kindergarten entry profiles, overview and initial statewide results, Fall 2003.* Sacramento, CA.

California Children and Families Commission. (n.d.). *First 5 school readiness initiative: Kindergarten entry profiles, overview and preliminary statewide results, Fall 2004.* Sacramento, CA.

Los Angeles Unified School District

Maddahian, E. (1998). *School Readiness Language Development Program evaluation: A student outcomes study.* Los Angeles, CA: LA Unified School District.

Connecticut

School Readiness Program

Bond, J. (2000). *Interim report: The evaluation of Connecticut's school readiness program cohorts 1 and 2 through spring 2000.* Hartford, CT: Department of Social Services and Department of Education.

Delaware

Early Childhood Programs

McCormick-Gamel & Amsden (2002). *Investing in better outcomes: The Delaware early childhood longitudinal study.* Newark, DE: Center for Disabilities Studies.

District of Columbia

Public Preschool Program

Marcon, R. A. (2000). *Educational transitions in early childhood, middle childhood and early adolescence: Head Start vs public school prekindergarten graduates.* Paper presented at the Fifth National Head Start Conference, Washington, DC.

Florida

Miami-Dade County Prekindergarten Program

Levitt, J. (2002). *First interim report prekindergarten longitudinal study 1993–2007, grade 5 1999–2000 overall school outcome analysis.* Miami, FL: Miami-Dade Public Office of Evaluation and Research.

Georgia

Georgia Universal Prekindergarten Program

Andrew Young School of Policy Studies. (2000). *Prekindergarten longitudinal study: Findings from the 1998–99 school year.* Atlanta, GA: Georgia State University: Author.

Henderson, L., Basile, K., & Henry, G. (1999). *Pre-kindergarten longitudinal study 1997–98 school year annual report.* Atlanta, GA: Georgia State University, Applied Research Center.

Henry, G. T., Gordon, C. S., Henderson, L.W., & Ponder, B. D. (2003). *Georgia Pre-K longitudinal study: Final report 1996-2001.* Atlanta, GA: Georgia

State University, Andrew Young School of Policy Studies.

Henry, G. T., Gordon, C. S., Mashburn, A., & Ponder, B. D. (2001). *Pre-K longitudinal study: Findings from the 1999–2000 school year.* Atlanta, GA: Georgia State University, Applied Research Center.

Henry, G. T., Ponder, B., Rickman, D., Mashburn, A., Henderson, L.W., & Gordon, C. (2004). *An evaluation of the implementation of Georgia's pre-K program: Report of the findings from the Georgia early childhood study (2002–03).* Atlanta, GA: Georgia State University, Andrew Young School of Policy Studies.

Henry, G. T., Rickman, D., Ponder, B., Henderson, L., Mashburn, A., & Gordon, C. (2005). *The Georgia early childhood study, 2001–2004.* Atlanta, GA: Georgia State University, Andrew Young School of Policy Studies.

Georgia Summer Readiness Pilot Program

Ponder, Rickman, & Henry, G. (2004). *Evaluation of the summer readiness pilot program.* Atlanta, GA: Georgia State University, Andrew Young School of Policy Studies.

Illinois

Chicago Child-Parent Centers

Clements, M. A., Reynolds, A. J., & Hickey, E. (2004). Site-level predictors of children's school and social competence in the Chicago Child-Parent Centers. *Early Childhood Research Quarterly, 19*, 272–296.

Reynolds, A. J. (1997). The Chicago Child-Parent Centers: A longitudinal study of extended early childhood intervention. *Institute for Research on Poverty Discussion Paper No. 1126–97.*

Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2001). Long-term effects of an early childhood intervention on educational achievement and juvenile arrest: A 15-year follow-up of low-income children in public schools. *Journal of the American Medical Association, 285*(18), 2339–2346.

Temple, J. A., Reynolds, A. J., & Meidel, W. T. (2000). Can early intervention prevent high school dropout? Evidence from the Chicago Child-Parent Centers. *Urban Education, 35*(1), 31–56

Illinois Prekindergarten Program for At-Risk Children

Illinois State Board of Education (2001). *Illinois pre-kindergarten program for children at risk of academic failure: FY 2000 evaluation report.* Chicago, IL.

Iowa

Shared Visions Programs for At-Risk Four-Year-Olds

Zan, B., & Edmiaston, R. (2000). *Evaluation of Shared Visions Programs for At-Risk Four-Year-Olds: Technical report, phase three.* Des Moines, IA: Child Development Coordinating Council Regents' Center Evaluation.

Kansas

Four-Year-Old At-Risk Program

Martinez, S. (2000). *Kansas state department of education study of programs for at-risk four year olds: The work sampling system developmental checklist section.* Topeka, KS: Department of Education, Department of Planning and Research.

Martinez, S. (2002). *Four-Year-Old At-Risk Program: Final evaluation.* Retrieved from www.ksde.org/pre/atriskpreschool.doc

Kentucky

Preschool Program

Hemmeter, M. L. (2001). Kentucky Preschool Program. *2000 review of research on the Kentucky Educational Reform Act*. Retrieved from www.kier.org/preschool.html

Louisiana

LA 4

Louisiana Department of Education. (2002). *LA4 prekindergarten program evaluation: Pilot year January–June 2002.* Baton Rouge, LA: Author.

Louisiana Department of Education. (2003). *LA 4 prekindergarten program evaluation 2002–2003.* Baton Rouge, LA: Author.

Starting Points

Louisiana Department of Education. (2001). *Evaluation of the Starting Points program: A follow-up study (children of years 1997–98 through 2000–2001).* Baton Rouge, LA: Author.

LA 4 and Starting Points

Louisiana Department of Education. (2004). *LA 4 and Starting Points prekindergarten evaluation 2003–2004.* Baton Rouge, LA: Author.

Louisiana Department of Education. (2005). *LA 4 and Starting Points prekindergarten program evaluation 2004–2005.* Baton Rouge, LA: Author.

Maryland

Judith Centers

eQuotient, Inc. (2002). Alleghany County Judy Center evaluation: January 2001–June 2002. Cumberland, MD: Author.

Overlook Judy Center Partnership. (2005). *Overlook Judy Center partnership FY 2005 evaluation report.* Garrett County, MD: Author.

Michigan

Full Day Preschool Program

Jurkiewick, T., Schweinhart, L., & Xiang, Z. (2004). *Realizing the potential: Final report of the Michigan Full Day Preschool Program comparison study.* Ypsilanti, MI: High/Scope Educational Research Foundation.

School Readiness Program

Lamy, C., Barnett, W., & Kwanghee, J. (2005). *The effects of the Michigan School Readiness Program on young children's abilities at kindergarten entry.* Rutgers, NJ: National Institute for Early Education Research.

Xiang, Z., & Schweinhart, L. (2002). *Effects five years later: The Michigan School Readiness Program evaluation through age 10.* Ypsilanti, MI: High/Scope Educational Research Foundation.

Xiang, Z., Schweinhart, L., Hohmann, C., Smith, C., Storer, E., & Oden, S. (2000). *Points of light: Third year report of the Michigan school readiness evaluation*. Ypsilanti, MI: High/Scope Educational Research Foundation.

Saginaw Michigan Public Schools

Kurecka, P., & Claus, R. N. (2000, July). *Pre-kindergarten program product evaluation report, 1999–2000: An approved report of the Department of Evaluation, Testing, and Research, school district of Saginaw, Michigan*. Saginaw, MI: Saginaw Public Schools Department of Evaluation Services.

Minnesota

School Readiness 1999/2000

Mueller, M. R. (2001). *Minnesota's school readiness 1999–2000: Immediate outcomes demonstrated by children participating in school readiness.* Minneapolis, MN: Mueller Associates.

Missouri

HB 1519 Early Childhood Project

Thornburg, K., Fuger, K., Mayfield, W., & Mathews, W. (2003). *House bill 1519 early childhood project:*

Executive summary and policy recommendations. Columbia, MO: University of Missouri, Center for Family Policy and Research and Institute for Human Development.

Nebraska

Early Childhood Education Grant Program

Jackson, B., & St. Clair, L. (2004). *Nebraska early childhood education grant program: Annual evaluation report.* Lincoln, NB: Department of Education and Child Development, University of Nebraska Medical Center.

Nevada

Early Childhood Education Program

Leitner, D. (2003). *Senate bill 585: Nevada Early Childhood Education (ECE) Program—Final evaluation report.* 2002–03 evaluation report. Carson City, NV: Nevada Department of Education.

New Jersey

Abbott Preschool Program

Frede, E., Lamy, C., Seplocha, H., Strasser, J., Juncker, J., & Wolock, E. (2004). *A rising tide: Classroom quality and language skills in the Abbott preschool program: Year 2 preliminary update 2003–2004.* Trenton, NJ: New Jersey Department of Education.

Lamy, C., Barnett, S., & Kwanghee, J. (2005a). *The effects of New Jersey's Abbott preschool program on young children's school readiness.* New Brunswick, NJ: Rutgers University: National Institute for Early Education Research.

Lamy, C., Frede, E., Seplocha, H., Saigeetha, J., Strasser. J., Juncker, J., Ferrar, H., & Wiley, L. (2004). *Inch by inch, row by row, gonna make this garden grow: Classroom quality and language skills in the Abbott preschool program.* Trenton, NJ: State Department of Education.

Lamy, C., Frede, E., Seplocha, H., Strasser, J., Saigeetha, J., Juncker, J., & Wolock, E. (2005b). *Giant steps for the littlest children: Progress in the sixth year of the Abbott preschool program: Year 3 initial update 2004–2005.* Trenton, NJ: State Department of Education.

New York

Rochester Early Childhood Assessment Partnership

Gramiak, W., Hightower, A., & Baker, A. (2005). *Rochester Early Childhood Assessment Partnership 2004–2005.* (Technical report #T05-002). Rochester, NY: Children's Research Institute.

Gramiak, W., Hightower, A., Brugger, L., Montes, G., Greenberg, S., & MacGowan, A. (2004). *Rochester Early Childhood Assessment Partnership 2003–2004 seventh annual report.* (Technical Report #T04-007). Rochester, NY: Children's Institute.

Montes, G., Hightower, A., Brugger, L., Moustafa, E., Greenberg, S., & MacGowan, A. (2002). *Rochester Early Childhood Assessment Partnership 2001–2002 annual report.* (Technical Report #T02-020). Rochester, NY: Children's Institute.

Montes, G., Hightower, A., Brugger, L., Moustafa, E., Greenberg, S., Gramlak, W., & MacGowan, A. (2003). *Rochester Early Childhood Assessment Partnership 2002–2003 annual report.* (Technical Report # T03-004). Rochester, NY: Children's Institute.

Montes, G., & Hoffman, D. (2004). *Who benefits from high quality pre-kindergarten? Investigating ethnicity/race differences in the RECAP 2002 sample.* (Technical Report # T04-004). Rochester, NY: Children's Institute.

North Carolina

More at Four

Peisner-Feinberg, E., & Maris, C. (2003). *Evaluation of the North Carolina More at Four Pre-kindergarten*

Program year 2 (July 2002–June 2003). Chapel Hill, NC: University of North Carolina, Frank Porter Graham Child Development Institute.

Peisner-Feinberg, E., & Maris, C. (2005). *Evaluation of the North Carolina More at Four Pre-kindergarten Program year 3 (July 2003–June 2004).* Chapel Hill, NC: University of North Carolina, Frank Porter Graham Child Development Institute.

Peisner-Feinberg, E., Elander, K., & Maris, C. (2006). *Evaluation of the North Carolina More at Four Pre-kindergarten Program year 4 (July 2004–June 2005).* Chapel Hill, NC: University of North Carolina, Frank Porter Graham Child Development Institute.

Smart Start

Bryant, D., Bernier, K., Taylor, K., & Maxwell, K. (1998). *The effects of Smart Start childcare on kindergarten entry skills.* Chapel Hill, NC: University of North Carolina, Frank Porter Graham Child Development Institute, and Orange County Partnership for Young Children.

Bryant, D., Maxwell, K., Taylor, K., Poe, M., Peisner-Feinberg, E., & Bernier, K. (2003). *Smart Start and preschool childcare quality in North Carolina: Change over time and relation to children's readiness.* Chapel Hill, NC: University of North Carolina, Frank Porter Graham Child Development Institute.

Maxwell, K., Bryant, D., & Miller-Johnson, S. (1999). *A six-county study of the effects of Smart Start child care on kindergarten entry skills*. Chapel Hill, NC: University of North Carolina, Frank Porter Graham Child Development Institute.

Charlotte-Mecklenburg, North Carolina—Bright Beginnings

Smith, E., Pellin, B., & Agruso, S. (2003). *Bright Beginnings: An effective literacy-focused preK program for educationally disadvantaged four-year-old children.* Alexandria, VA: Educational Research Service.

Ohio

Head Start

Cogswell, S. H., Lochtefeld, S. S., Skaggs, A. V., Walker, J. P. (1998*). Head Start's impact on school readiness in Ohio: A case study of kindergarten students*. Columbus, OH: Legislative Office of Education Oversight. Retrieved from www.loeo.state.oh.us

Columbiana County Head Start. (2004). *Community action agency of Columbiana County, Inc., Head Start child outcomes and evaluation report, 2003–2004*. Lisbon, OH: Author.

Oklahoma

Prekindergarten Program

Gromley, W., & Gayer, T. (2003). *Promoting school readiness in Oklahoma: An evaluation of Tulsa's prekindergarten program*. (Working Paper # 1). Washington, DC: Georgetown University, Public Policy Institute.

Gromley, W., Gayer, T., Phillips, D., & Dawson, B. (2005). Effects of universal pre-K on cognitive development. *Developmental Psychology, 41*(6). 872–884.

Gromley, W., Gayer, T., Phillips, D., & Dawson, B. (2004). *The effects of Oklahoma's early childhood four year old program on young children's school readiness*. Washington, DC.: Georgetown University, Public Policy Institute.

Pennsylvania

Pittsburgh Early Childhood Initiative

Bagnato, S. (2002). *Quality early learning—key to school success: A first-phase 3-year program evaluation research report for Pittsburgh's early childhood initiative (ECI).* Pittsburg, PA: SPECS Evaluation Team, Early Childhood Partnerships, Children's Hospital of Pittsburgh.

Bagnato, S., Suen, H., Brickley, D., Smith-Jones, J., & Dettore, E. (2002). Child development impact of Pittsburgh's early childhood initiative in high risk communities: First phase authentic evaluation research. *Early Childhood Research Quarterly, 17*(4), 559–580.

South Carolina

Greenville County School District

Coleman, B., & McCreary, J. (2003). *Effectiveness of the 4K program*. Greenville, SC: Greenville County School District.

Child Development Program for Four-Year-Olds

Lamy, C., Barnett, W., & Kwanghee, J. (2005). *The effects of South Carolina's early education programs on young children's school readiness.* New Brunswick, NJ: Rutgers University, National Institute for Early Education Research.

South Carolina State Department of Education. (2002). *What is the penny buying for South Carolina? Child Development Programs for Four-Year-Olds: Student and program characteristics, longitudinal study of academic achievement and current parent perceptions.* Columbia, SC: Author.

South Carolina State Department of Education. (2004). *What is the penny buying for South Carolina? Twentieth annual reporting on the South Carolina education improvement act of 1984: Child Development Programs for Four-Year-Olds: Longitudinal studies of later academic achievement, 1995–96 through 1999–2000 and 2000–01 through 2001–02.* Columbia, SC: Author.

Yao, W., Snyder, C., Burnett, D., Lindsay, S., & Tenenbaum, I. (2000). *A longitudinal research report on the early childhood development program: The half day Child Development Program for Four-Year-Olds, 1997–98.* Columbia, SC: South Carolina State Department of Education.

Texas

Austin Independent School District Prekindergarten Expansion Grant Program

Curry, J. (2000). *Prekindergarten Expansion Grant evaluation, 1999–2000.* Austin, TX: Austin Independent School District, Office of Program Evaluation.

Curry, J. (2001). *Prekindergarten Expansion Grant evaluation, 2000–2001.* Austin, TX: Austin Independent School District, Office of Program Evaluation.

Curry, J. (2002). *Prekindergarten Expansion Grant evaluation, 2001–2002.* Austin, TX: Austin Independent School District, Office of Program Evaluation.

Curry, J. (2003). *Prekindergarten expansion grant evaluation, 2002–2003.* Austin, TX: Austin Independent School District, Office of Program Evaluation.

Curry, J. (2004). *Prekindergarten Expansion Grant evaluation, 2003–2004.* Austin, TX: Austin Independent School District, Office of Program Evaluation.

Curry, J. (2005). *Prekindergarten Expansion Grant evaluation, 2004–2005.* Austin, TX: Austin Independent School District, Department of Program Evaluation.

Dallas Independent School District Prekindergarten Expansion Grant Program

Dallas Independent School District. (2004). *Evaluation of the prekindergarten expansion grant: 2003–2004.* Report No. REIS04-171-2. Dallas, TX: Author.

Dallas Independent School District. (2005). *Final evaluation of the 2004–2005 prekindergarten expansion grant:* Report No. REIS05-171-2. Dallas, TX: Author.

Ft. Worth Independent School District

Mindel. (2005). *Child Care Associates: Annual evaluation report 2003–2004*. Ft. Worth, TX: M & D Research and Evaluation.

Language Enrichment Activities Program

Dallas Independent School District. (2005). *Final evaluation of 2004–2005 Language Enrichment Activities Program (LEAP).* Report No. REIS05-172-2. Dallas, TX: Author.

Washington

Early Childhood Education Assistance Program (ECEAP)

Northwest Regional Educational Laboratory, Child and Family Program. (1998). *Early Childhood Education and Assistance Program: An investment in children and families: Year 7 longitudinal study report*. Portland, OR: Author.

Northwest Regional Educational Laboratory, Child and Family Program. (2000). *Early Childhood Education and Assistance Program: An investment in children and families. Years 9 & 10 longitudinal study report.* Portland, OR: Author.

West Virginia

Early Education Program

Lamy, C., Barnett, W., & Kwanghee, J. (2005). *The effects of West Virginia's Early Education Program on young children's school readiness.* New Brunswick, NJ: Rutgers University, National Institute for Early Education Research.

APPENDIX D
DESCRIPTIONS OF MOST COMMONLY USED INSTRUMENTS

TABLE D1
Abstracts of key instruments

1. Basic School Skills Inventory, Third Edition	
Author(s)	Hammill, D. D., & Leigh, J. E.
Publisher	Pro-Ed, Inc. Publishing
Web site	www.proedinc.com
Intended purpose	Aiding teachers in making decisions about programming and instruction (Hammill, Leigh, Pearson, & Maddox, 1998)
Age range	4 years to 6 years 11 months
Domains measured	Basic knowledge, language/communication skills, literacy skills, math/pre-math skills, classroom behavior
Reliability	*Internal consistency* was examined using coefficient alpha, and correlations were provided by subtest and age. Of 30 correlations all but 2 were greater than .85. The composite was .98. Stability reliability was examined with a two-week interval for 49 kindergarten through third-grade regular education students. Correlations ranged from .96 to .99.
Validity	*Content validity* was suggested by the test development process and the test format selected. Items on the test were based on teacher descriptions of "ready" and "not ready" children. Items were field tested several times. *Criterion-related validity* was based on a comparison with the Rhode Island Test of Language Structures and the Expressive One-Word Picture Vocabulary Test–R. Correlations ranged from .37 to .87. Additional studies were suggested. For *construct validity*, means increased with age for all subtests except behavior. Results from Classroom Behavior and Daily Living Skills were compared with the Hawaii Early Learning Profile subtests Self-help, Fine and Gross Motor Skills, and Social/Emotional. Resulting correlations were .36–.64. For academic areas correlations were .35–.71 when compared with the Brigance Preschool Screen–R. *Concurrent validity* was established by correlating subtests with other language development, self-help, social, and general knowledge tests. Results were based on a sample of 42 preschool children who received special services. Discriminant validity was based on a similar sample.
Languages available	English
Administration time	4–8 minutes
Training recommended	Formal training in assessment, familiarity with preschool classroom skills and behavior/social/emotional testing
States that used this instrument	Pennsylvania

2. Battelle Developmental Inventory	
Author(s)	Newborg, J., Stock, J. R., & Wnek, L.
Publisher	Riverside Publishing
Web site	www.riverpub.com
Intended purpose	Screening, tracking child outcomes
Age range	Infant through 7 years 11 months
Domains measured	Cognitive development, social/emotional development, language/communication skills, child health/physical development
Reliability	Reliabilities meet or exceed traditional standards for excellence at the subdomain, domain, and full test composite levels.
Validity	*Concurrent and criterion validity* were obtained using the original Battelle Developmental Inventory; the Bayley Scales of Infant Development–Second Edition; Woodcock-Johnson III; Denver Developmental Screening Test–Second Edition; Preschool Language Scale–Fourth Edition; Vineland Social-Emotional Early Childhood Scales; and Wechsler Preschool and Primary Scale of Intelligence–Third Edition.
Languages available	English and Spanish
Administration time	Complete assessment 1–2 hours; screening 10–30 minutes
Training recommended	The manual (2005) suggests that the instrument may be administered by teachers, special educators, infant interventionists, psychologists, speech and language pathologists, diagnosticians, health professionals, and other related service providers. There are further recommendations that examiners have appropriate training and experience in administering the instrument, as well as knowledge and familiarity with children within the age range being assessed.
States that used this instrument	Kentucky

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

3. Bracken Basic Concepts Scale–Revised (BBCS–R)	
Author(s)	Bracken, B. A.
Publisher	Harcourt, Brace and Co.
Web site	www.harcourt.com
Intended purpose	Assessing children's concept development and determining how familiar children are with concepts that parents, preschool teachers, and kindergarten teachers teach children to prepare them for formal education (as described by the author).
Age range	2 years 6 months to 8 years
Domains measured	Cognitive development, language/communication skills, math/pre-math skills, social/emotional development, school readiness (composite of first six subtests)
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. Analyses were conducted using the School Readiness Composite (SRC), subtests 7–11, 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 (users manual). *Test-retest reliability* analyses were conducted using the SRC and individual tests 7–11. The test-retest reliability of the SRC was .88. The test-retest reliabilities of subtests 7–11 ranged from .78 to .94 (users manual).
Validity	*Internal validity* correlations were calculated for each age group (2–7 years), as well as for the full sample, between scores on the SRC subtests 7–11, and the full battery. Intercorrelations among the SRC and scores on subtests 7–11 for the full sample ranged from .58 to .72. In the full sample, intercorrelations between subtests 7–11 and total test scores ranged from .79 to .87. The intercorrelation 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. *Concurrent validity.* According to the users manual, a number of studies have indicated that children's scores on the BBCS–R are 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. *Predictive and discriminant validity.* Information on predictive and discriminant validity indicates that scores on the BBCS–R are associated with later performance in school and that children with known language or developmental delays differ in their BBCS–R performance. Substantial percentages of children were not correctly identified on the basis of their BBCS–R scores (users manual).
Languages available	English and Spanish
Administration time	Untimed assessment, typically takes 30 minutes
Training recommended	Training in psychological testing interpretation. It is recommended that the user have graduate training in measurement, guidance, individual psychological assessment, or special appraisal methods appropriate to a particular test.
States that used this instrument	Connecticut

4. California Preschool Social Competency Scale	
Author(s)	Levine, S., Elzey, F. F., & Lewis, M.
Publisher	Formerly published by Consulting Psychologists Press
Web site	Not available
Intended purpose	Assessing children's social adjustment in the classroom
Age range	Preschool-age children
Domains measured	Social/emotional development, school adjustment/attitude
Reliability	The scale has been shown to have acceptable interrater (.75–.86) and split-half (.90–.98) reliability (Levine, California Preschool Competency Scale, 1969).
Validity	Not available
Languages available	English
Administration time	5–10 minutes
Training recommended	No training recommendations
States that used this instrument	Ohio

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

5. Child Observation Record (COR)	
Author(s)	High/Scope Educational Research Foundation
Publisher	High/Scope Educational Research Foundation
Web site	www.highscope.org
Intended purpose	Providing data for teachers to use in planning instruction. The authors also state that data from the COR can be used to document changes in children's progress over time, evaluate a curriculum, and evaluate a classroom or program.
Age range	2½ years to 6 years
Domains measured	Initiative, social relations, creative representation, movement/music, language/literacy skills, math/pre-math skills
Reliability	Data were collected from 160 children in spring 2002 and from 233 children in fall 2002. The reliability of the total COR scale was high (.94 in the spring and .91 in the fall samples). The *internal consistency* of the subscales was lower but acceptable (.75–.88). *Interrater reliability* was assessed by having 10 pairs of teachers and assistant teachers rate the same 41 children. Interrater reliability was .73 for the total COR score and .69–.79 for the subscales.
Validity	A confirmatory factor analysis indicated that the scale exhibited four factors that roughly correspond to the subscales but combine two of the categories with other categories. The categories identified were mathematics and science, language and literacy, initiative/social relations, and creative representation/ movement and music. The authors found low to moderate correlations between children's COR rating and their score on the Cognitive Skills Assessment Battery (N= 28) and weak but significant correlations between children's age and their score on the COR (N = 233).
Languages available	English
Administration time	No specified time; observations are ongoing
Training recommended	It is recommended that administrators attend a two-day COR training.
States that used this instrument	Iowa, Michigan, Nebraska, New York, and North Carolina

6. Color Bears and Counting Bears	
Author(s)	Family and Child Experiences Survey (FACES) Research Team, measure modified from the Color Concepts and Number Concepts tasks in Mason & Stewart (1989).
Publisher	Unpublished
Web site	www.acf.hhs.gov/programs/opre/hs/faces/instruments/child_instru02/language_color.pdf
Intended purpose	Assessing knowledge of colors and counting ability
Age range	3–5 years
Domains measured	Early literacy and numeracy
Reliability	*Internal consistency* (Cronbach's Alpha) Color Names, correlative coefficient of .94
Validity	*Predictive validity:* The correlations between Color Names and Counting scores at the end of the spring 1998 Head Start year and the Early Childhood Longitudinal Study–K (ECLS–K) Reading scale scores at the end of the spring 1999 kindergarten year were .39 for Color Names and .40 for Counting. The correlations between Color Names and Counting scores at the end of the spring 1998 Head Start year and ECLS–K General Knowledge scale scores at the end of the spring 1999 kindergarten year were .38 for Color Names and .36 for Counting. In multivariate regression analyses with the scale scores from the entire FACES battery at the end of the Head Start year predicting ECLS–K Reading scores at the end of the kindergarten year, counting (beta = .12) tasks were a significant predictor in the model.
Languages available	English
Administration time	5 minutes
Training recommended	Paraprofessionals can be trained in about 15 minutes.
States that used this instrument	Georgia, North Carolina

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

7. Creative Curriculum Developmental Continuum for Ages 3 – 5 Assessment	
Author(s)	Trister-Dodge, D., Colker, L., & Heroman, C.
Publisher	Teaching Strategies, Inc.
Web site	www.teachingstrategies.com
Intended purpose	Informing instruction, tracking child outcomes, conducting research
Age range	3 years through kindergarten
Domains measured	Cognitive development, general knowledge and awareness, social/emotional development, health/physical development, language/communication skills
Reliability	Relatively high degree of reliability (all alphas above .92). No information on overall reliability of the entire Developmental Continuum or of the subscales.
Validity	A factor analysis was used to examine the Creative Curriculum's *construct validity.* Results indicate that data fell into four factors, suggesting that four separate constructs are being measured. Results support the *content validity* of the Developmental Continuum, particularly in the domains of social/emotional and health/physical development; however, the Early Literacy (reading and writing) subscale may be collecting data that are related more closely to children's overall cognitive development than specifically to early literacy skills.
Languages available	English and Spanish
Administration time	Time is not specified
Training recommended	Publisher offers a self-paced training module series that offers guidance on how to conduct authentic assessments and how to use the Creative Curriculum Assessment tool. Publisher also offers other training opportunities. Web and software training are also available.
States that used this instrument	Louisiana, Nebraska

8. Comprehensive Test of Phonological Processing (CTOPP)	
Author(s)	Wagner, R., Torgeson, J., & Rashotte, C.
Publisher	Pro-Ed, Inc.
Web site	www.proedinc.com
Intended purpose	Screening, tracking child outcomes, conducting research
Age range	5 years through 24 years 11 months
Domains measured	Language/communication skills
Reliability	*Test-retest coefficients* range from .70 to .92. The coefficients showed a mean of .82 for ages 5–7, .80 for ages 8–17, and .79 for ages 18–24. Internal consistency ranged from .70 (for 7-year-olds on the Rapid Letter Naming Test) to .96 (for 12-year-olds on the Rapid Digit Naming Test), with a mean of .87. Average internal consistency exceeded .80. Coefficient alpha was computed with a range of demographic groups on the CTOPP subtests. Results ranged from .68 to .97, with a mean of .89. Interrater reliability was .98 according to the test manual. Coefficients from all reliability studies suggest limited error and good reliability (Spies & Plake, 2005).
Validity	*Content validity*, *criterion-related validity*, and *construct validity* are reported. Item bias was examined using delta scores resulting in correlation coefficients with a mean of .98. Criterion-related validity was examined based on correlations between Comprehensive Test of Phonological Processing composite scores and subtests of the Woodcock Reading Mastery Test–Revised (WRMT–R). Coefficients were .71, .42, and .66 one year after kindergarten and .80, .52, and .70 one year after first grade. The instrument was also validated separately with a group of students in kindergarten through fifth grade.
Languages available	English
Administration time	30 minutes
Training recommended	There are no specific training recommendations. The test manual provides explicit instructions. Examiners may contact the publisher if they need clarification on any administration issues. However, extensive training in assessment with an emphasis on phonological ability testing, test statistics scoring, and interpretation is recommended (Spies & Plake, 2005).
States that used this instrument	Georgia

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

9. Developmental Observation Checklist System	
Author(s)	Hresko, W., Miguel, S. , Sherberon, R., & Burton, S.
Publisher	Pro-Ed, Inc.
Web site	www.proedinc.com
Intended purpose	Screening, tracking child outcomes
Age range	Birth through first grade
Domains measured	Language/communication, cognitive development, social/emotional development, motor movement
Reliability	For *internal consistency* the Cronbach alphas for infants through 3-year-olds ranged from .80 to the mid to high .90s. *Test-retest reliability*, using a 14–21 day interval for ages 2–3 years, ranged from .85 to .94. *Interrater reliability* ranged from. 91 to .94. (Information retrieved from www.acf.hhs.gov/programs/opre/ehs/perf_measures/reports/resources_measuring/resources_for_measuring.pdf)
Validity	*Construct validity* is supported through correlations with age and group differentiation relating test items to total test scores, component intercorrelations, and cognitive aptitude. Delta values confirm the nondiscriminatory basis of the items with respect to gender and race. Substantial *content validity* and *criterion-related validity* are offered.
Languages available	English
Administration time	30 minutes to complete and 15–20 minutes to score all three checklists
Training recommended	Examiners should have some training in administering and interpreting assessment instruments. The instrument can be completed by a parent with a fourth-grade reading level.
States that used this instrument	Pennsylvania

10. Developing Skills Checklist (DSC)	
Author(s)	CTB-Macmillan-McGraw Hill
Publisher	CTB-McGraw Hill
Web site	www.ctb.com
Intended purpose	Planning instruction
Age range	4–6 years
Domains measured	Language/communication skills, visual, auditory, math/pre-math skills, memory, print and writing, social/emotional development, fine and gross motor movement
Reliability	The DSC was normed on a sample of 3,985 individuals. The sample was relatively representative of the U.S. population in terms of demographics such as gender and race, and the sample was stratified based on school size, community socioeconomic status, and geographic region. Internal consistency was reflected by KR 20s, a statistical measurement of reliability. KR 20s were between .81 and .95 for all scales except the visual scale, which had a mean of .69. No *test-retest reliability* information was available.
Validity	*Construct validity,* based on comparison of the DSC with the Early School Assessment, was weak. No *predictive validity* information was available.
Languages available	English and Spanish
Administration time	Not a timed test, but allow 10–15 minutes for each of the three testing sessions
Training recommended	There are no specific training recommendations. The test manual gives explicit instructions. Examiners may contact the publisher if they need clarification on any administration issues.
States that used this instrument	Louisiana, Texas

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

11. Developmental Indicators for the Assessment of Learning–Third Edition (DIAL–3)	
Author(s)	Mordell-Czudnowski, C., & Goldenberg, C.
Publisher	American Guidance Systems
Web site	http://www.agsnet.com/
Intended purpose	Screening, informing instruction
Age range	3 years through second grade
Domains measured	Social/emotional development, health/physical development, language/communication skills, self-help, concepts
Reliability	*Internal consistency*. Coefficient alphas for the total scale range from .77 to .90, with the lowest alpha in the older age range (6 years 6 months to 6 years 11 months). The median alpha is .87. Alphas for Speed DIAL (a brief screening alternative) are lower and range from .69 (again in the older age range) to .85, with a median of .80. *Test-retest reliability* coefficients for the total DIAL-3 scale are .86 for younger children and .82 for older children. Coefficients for the subscales range from .65 to .86, and the Speed DIAL coefficients were .80 and .77.
Validity	To assess validity, the developers compared children's scores on the Early Screening Profiles, the Battelle Screening Test, the Bracken Screening Test, the Brigance Preschool Screen, the Differential Ability Scales, the Peabody Picture Vocabulary Tests–III, and the Social Skills Rating System. The resulting correlation coefficients ranged from .17 to .79, with most in the "modest" range.
Languages available	English and Spanish
Administration time	30 minutes
Training recommended	The manual recommends 4 hours of training. Administrators should demonstrate competence in administering, scoring, and interpreting the DIAL–3 area in English or Spanish and the ability to relate to young children of any linguistic or cultural background.
States that used this instrument	South Carolina, Texas

12. Expressive One-Word Picture Vocabulary Test (EOWPVT)	
Author(s)	Brownell, R.
Publisher	Academic Therapy Publications
Web site	www.academictherapy.com
Intended purpose	Screening, monitoring growth, and evaluating program effectiveness
Age range	2 years through 18 years 11 months
Domains measured	Language/communication skills
Reliability	*Internal consistency.* Coefficient alphas range from .93 to .98, with a median of .96 across age groups. *Split-half reliability* coefficients range from .96 to .99, with a median of .98. *Test-retest reliability* correlations on a sample of 226 children ranged from .87 to .97 for different age groups, with a coefficient of .90 for the full sample. *Interrater reliability.* Twenty children were each tested by two different examiners, and then the protocols were scored by a single examiner. The corrected correlation between scores from the two protocols was .93.
Validity	*Concurrent validity.* The EOWPVT correlates with other tests of vocabulary (including the Peabody Picture Vocabulary Test–III, the Weschler Intelligence Scale for Children–III Vocabulary, and the Stanford-Binet Intelligence Test) with a range of .67 to .90 and a median of .79.
Languages available	English and Spanish
Administration time	10–15 minutes
Training recommended	Usually administered by someone with a relevant background (such as a speech pathologist or a psychologist). However, with training and supervision, it can be administered by someone without such a background.
States that used this instrument	Texas

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

13. Get it! Got it! Go!	
Author(s)	Early Childhood Research Institute on Measuring Growth and Development
Publisher	University of Minnesota, College of Education and Human Development
Web site	http://ggg.umn.edu
Intended purpose	Monitoring change
Age range	30–66 months
Domains measured	Literacy skills
Reliability	For Picture Naming the instrument has one-month *alternate form reliability coefficients* that range from .44 to .78. For alliteration *test-retest reliability* over three weeks ranges from.46 to .80, $p < .01$, for a sample of 42 preschool-aged children. For Rhyming *test-retest reliability* over three weeks ranges from .83 to .89, $p < .01$, for a sample of 42 preschoolers.
Validity	Picture Naming. When compared with the Peabody Picture Vocabulary Test–III and the Preschool Language Scale–Third Edition for concurrent validity, correlation coefficients range from .47 to .69. Picture Naming also correlates with chronological age ($r = .41$ in a longitudinal study and .60 in a cross-sectional study), including children from a variety of backgrounds ($r = .63$), children enrolled in Head Start ($r = .32$), and children with disabilities ($r = .48$). Alliteration. The EOWPVT correlates with the Peabody Picture Vocabulary Test–III ($r = .57$), Concepts about Print ($r = .55$), letter identification ($r = .74$), and Test of Phonological Awareness ($r = .75$). It also correlates with chronological age ($r = .61$). Rhyming. Correlates with Peabody Picture Vocabulary Test–III ($r = .56$), Concepts about Print ($r = .54$), letter identification ($r = .59$), Test of Phonological Awareness ($r = .62$), and chronological age ($r = .44$).
Languages available	Spanish available for picture naming subtest only
Administration time	5 minutes per test
Training recommended	The assumption is that all individuals using this measure from the web site will have basic familiarity with and skill in administering standardized tests to young children. At a minimum evaluators should review the section of the web site titled "Why standardized administrations matter."
States that used this instrument	Texas

14. Get Ready to Read	
Author(s)	Whitehurst, G., & Lonigan, C.
Publisher	Pearson Early Learning
Web site	www.pearsonassessments.com
Intended purpose	Screening
Age range	4-year-olds
Domains measured	Literacy skills
Reliability	Coefficient alpha is .78, the split-half coefficient is .80, and the standard deviation is 4.31, using a range of 20 and mean correct answers of 9.14.
Validity	The instrument has good validity. The instrument correlates with the Developing Skills Checklist, with a measure of letter naming, a measure of language development, and a battery of phonological awareness tests.
Languages available	English and Spanish
Administration time	9.5 minutes
Training recommended	Standardized training is available, both initial and follow-up.
States that used this instrument	New Jersey

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

15. Learning Accomplishments Profile–Revised (LAP–R)	
Author(s)	Developed through a partnership between the Chapel Hill Training Outreach Project, Kaplan Early Learning Company, and Red-E-Set-Grow.
Publisher	Kaplan Press
Web site	www.kaplanco.com
Intended purpose	Informing instruction; also for tracking children's progress in the areas assessed when the assessment is conducted more than once.
Age range	36–72 months
Domains measured	Cognitive development, language/communication skills, self-help, motor movement, social skills
Reliability	*Test-retest reliability* coefficients ranged from .96 to .99, indicating a high degree of reliability between the results of the assessment in the two sessions for this subset of children. A study of *interrater reliability* found coefficients ranging from .81 to .98.
Validity	*Criterion validity* coefficients were calculated using the Battelle Developmental Inventory. Correlations were high for some of the domains (ranging from .70 to .92) and moderate for the remaining domains (.54–.69). The lower correlations were related primarily to the Communication Domain on the Battelle Developmental Inventory and the Personal/Social Domain on the LAP-3.
Languages available	English and Spanish
Administration time	Varies
Training recommended	A video on the LAP–R is available for purchase through Kaplan. Training is also available from the Chapel Hill Training Outreach Project and Kaplan.
States that used this instrument	Ohio

16. Oral and Written Language Scales (OWLS)	
Author(s)	Carrow-Woolfolk, E.
Publisher	American Guidance Systems/Pearson Assessments
Web site	www.pearsonassessments.com
Intended purpose	Screening, informing instruction, conducting research
Age range	3–21 years
Domains measured	Cognitive development, language/communication skills, literacy skills
Reliability	*Internal consistency.* Internal reliabilities included scores for listening comprehension (coefficient of .84), oral expression (.87), written expression (.87), oral composite (.91), and language composite (.93). *Test-retest reliability* scores included listening comprehension (.73–.80), oral expression (.77–.86), oral composite (.81–.89), written expression (.87–.88), and language composite (.87–.90). For *interrater reliability* the Oral Expression Scale and the Written Expression Scale of the OWLS each had a mean score for four age groups of .95.
Validity	The test manual reports correlations of OWLS scales with other measures of receptive and expressive language as well as with tests of cognitive ability and academic achievement. Also, the score profiles of seven clinical groups are compared with matched control samples.
Languages available	English
Administration time	15–40 minutes
Training recommended	Training in psychological assessment
States that used this instrument	Georgia

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

17. Peabody Picture Vocabulary Test–Third Edition (PPVT–III)	
Author(s)	Dunn, L., & Dunn, L.
Publisher	Pearson Assessments
Web site	www.pearsonassessments.com
Intended purpose	Screening, tracking child outcomes, conducting research
Age range	2 years 6 months through 17 years 11 months
Domains measured	Cognitive development, language/communication skills
Reliability	*Internal consistency*. Alpha coefficients: .92–.98 (median: .95). *Split-half reliability*: .86–.97 (median: .94). *Alternate-form reliability*: .88–.96 (median: .94). *Test-retest reliability*: .91–.94 (median: .92).
Validity	Average correlation of .69 with the OWLS Listening Comprehension Scale and .74 with the OWLS Oral Expression Scale. Correlations with measures of verbal ability are .91 (Weschler Intelligence Scale for Children Verbal Intelligence Quotient), .89 (Kaufman Adolescent and Adult Intelligence Test Crystallized IQ), and .81 (Kaufman Brief Intelligence Test Vocabulary).
Languages available	English and Spanish
Administration time	8–16 minutes
Training recommended	Formal training in psychometrics
States that used this instrument	Arizona, Georgia, Michigan, Missouri, New Jersey, North Carolina, Oklahoma, South Carolina, Texas, West Virginia

18. Preschool Comprehensive Test of Phonological Processing (Pre-CTOPP)	
Author(s)	Lonigan, C., Wagner, R., Torgesen, J., & Rashotte, C.
Publisher	Pro-Ed, Inc., Publishing. The Pre-CTOPP was an unpublished evaluation tool. The parts that will be published in August 2007 will be marketed under the name Test of Preschool Early Literacy (TOPEL) and will be available through Pro-Ed. The TOPEL will include the definitional vocabulary (expressive), phonological awareness, and print knowledge components from the Pre-CTOPP.
Web site	www.proedinc.com
Intended purpose	Assessing children's phonological awareness, phonological memory, and phonological access (Lonigan, 2006).
Age range	3–5 years
Domains measured	Early literacy skills
Reliability	Reliability coefficients for the Pre-CTOPP ranged from .74 to .88
Validity	Results for the validity studies on this instrument indicate that coefficients for phonological awareness ranged from .43 to. 62, coefficients for phonological memory ranged from .29 to .42, and those for phonological access ranged from .57 to .60.
Languages available	English
Administration time	30–45 minutes
Training recommended	Not available
States that used this instrument	Michigan, New Jersey, Oklahoma, South Carolina, West Virginia

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

19. Pre-Language Assessment Scales (Pre-LAS 2000)	
Author(s)	Duncan, S. E., & De Avila, E. A.
Publisher	CTB/McGraw Hill
Web site	http://www.ctb.com
Intended purpose	Measuring oral language proficiency and pre-literacy skills
Age range	4 years through first grade
Domains measured	Cognitive development, language/communication skills, literacy skills
Reliability	The test was designed using a sample of 50 Hispanic Head Start children ages 3–5. Researchers determined which words from an established sample were identified by at least 50 percent of the children. These words were used in creating the Rhyming and Alliteration subtests. The Cronbach's alpha for the test was .86–.90 for the English version and .66–.88 for the Spanish version.
Validity	Not available
Languages available	English and Spanish
Administration time	10–15 minutes to administer the oral language component and 5–10 minutes to administer the pre-literacy component
Training recommended	Testers need experience in test administration.
States that used this instrument	North Carolina

20. Preschool and Kindergarten Behavior Scales (PKBS)	
Author(s)	Merrell, K. W.
Publisher	Pro-Ed, Inc., Publishing
Web site	www.proedinc.com
Intended purpose	Screening, informing instruction, and conducting research
Age range	3–6 years
Domains measured	Social/emotional development
Reliability	*Internal consistency*. Coefficient alpha and *split-half reliability* coefficients were .96 and .94 for the total Social Skills Scale and .97 and .96 for the total Problem Behavior Scale; coefficients for subscales ranged from .81 to .97. *Test-retest reliability*. Coefficients at three weeks after administration were .58 for the Social Skills Scales and .86 for the Problem Behavior Scales; at a three-month interval the coefficients were .69 and .78. *Interrater reliability*. Coefficient was .38 for total Social Skills Scale (significant) but only .16 for the Problem Behavior Scale (not significant). Subscale coefficients ranged from .13 to .57 for teacher ratings and parent ratings. Again, the Problem Behavior Scale had lower coefficients.
Validity	*Content validity* was assessed by a panel of experts and by correlating items to total scale; correlations ranged from .35 to .80. *Construct validity* was assessed with factor analysis; factor loadings ranged from .43 to .81. Scores on PKBS were correlated with Social Skills Rating System, and moderate to strong coefficients of .76 were found for the Social Skills Scale and .83 for the Problem Behavior Scale. PKBS scores were also correlated with scores on the Matson Evaluation of Social Skills with Youngsters, and Social Skills scores were correlated at .84 with the Appropriate Social Skills Scale and Problem Behavior scores at .64 with the Inappropriate Assertiveness/Impulsivity subscale; correlations with the Conners Teacher Rating Scales were in the moderate to strong range; correlations with the School Social Behavior Scales (total scales) were correlated at .77.
Languages available	English and Spanish
Administration time	12 minutes
Training recommended	The instrument can be completed by anyone who knows the child well. Scoring and interpretation should be done by someone with knowledge of basic principles of educational and psychological testing. Training in understanding and assessing child behavioral and emotional problems is recommended.
States that used this instrument	Pennsylvania

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

21. Preschool Language Scale–Fourth Edition (PLS–4)	
Author(s)	Zimmerman, I., Steiner, V., & Pond, R.
Publisher	The Psychological Corporation
Web site	www.psychcorp.com
Intended purpose	Screening
Age range	Infant to 6 years 11 months
Domains measured	Expressive and receptive language
Reliability	*Test-retest reliability* is .82–.95 for the subscales and .90–.97 for the total language scale. *Internal consistency* is .66–.96. *Interrater reliability* is .99.
Validity	Research and testing were conducted on test content, response processes, internal structure, and relationships with other variables.
Languages available	English and Spanish
Administration time	20–40 minutes
Training recommended	Familiarity with the manual and with assessing young children is needed. Paraprofessionals can be trained to administer the instrument, but interpretation of results needs to be done by a clinician who has training and experience in diagnostic assessment.
States that used this instrument	Nevada

22. Receptive One-Word Picture Vocabulary Test (ROWPVT)	
Author(s)	Brownell, R.
Publisher	Academic Therapy Publications
Web site	www.academictherapy.com
Intended purpose	Assessing the ability to understand the spoken and written vocabulary of others.
Age range	Infant through fourth grade and beyond
Domains measured	Language/communication skills
Reliability	Alpha coefficients range from .93 to .98; split half = .98
Validity	Not available
Languages available	English and Spanish
Administration time	10–15 minutes
Training recommended	Specialized training is needed. Testers should have college-level work in psychology or counseling and work in testing or assessment, or they should be licensed in testing.
States that used this instrument	Texas

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

23. Social Skills Rating System (SSRS)	
Author(s)	Gresham, F. M., & Elliott, S. N.
Publisher	Pearson Assessments
Web site	http://ags.pearsonassessments.com
Intended purpose	Screening, informing instruction, tracking child outcomes, conducting research
Age range	3 years through kindergarten
Domains measured	Social/emotional development, academic competence
Reliability	*Internal consistency*. Teacher Form - Social Skills Total Scale: .94 (preschool), .94 (elementary), and .93 (secondary); Teacher Form - Problem Behaviors Scale: .82, .88, and .86 across age/grade levels. *Test-retest reliability*. Teacher ratings: .85 for Social Skills, .84 for Problem Behaviors, and .93 for Academic Competence.
Validity	*Content validity*. Established by basing items on review of literature and having experienced researchers nominate a pool of items; teachers, parents, and secondary school students then rated the importance of each of the social skills on the scale—indicated items rated as important. Teacher ratings on SSRS were correlated with ratings on the Social Behavior Assessment, yielding correlations on the total scale of .68 for Social Skills, .55 for Problem Behaviors, and –.67 for Academic Competence. SSRS was also compared with the Child Behavior Checklist Teacher Report Form, and the SSRS Problem Behaviors Total Score showed strong correlations (.81) with the total score and negative correlation (–.59) with the Academic Competence Scale. The SSRS Social Skills Scale showed a .70 correlation with the Harter Teacher Rating Scale total score, the Academic Competence Scale showed a .63 correlation with Harter, and the Problem Behaviors Scale showed a –.66 correlation with Harter.
Languages available	English and translated into Spanish for Family and Child Experiences Survey research team
Administration time	10–15 minutes per questionnaire
Training recommended	Training in psychological testing. Follow script and gesturing guidelines.
States that used this instrument	Connecticut, Georgia, Kentucky, Missouri, North Carolina

24. Story and Print Concepts	
Author(s)	Family and Child Experiences Survey Research Team, measure modified from the Story and Print Concepts task in Mason & Stewart (1989).
Publisher	Unpublished
Web site	www.acf.hhs.gov/programs/opre/hs/faces/instruments/child_instru02/language_story.pdf
Intended purpose	Assessing basic story concepts (such as comprehension of story content), print concepts (such as where the name of the book is written), and the mechanics of reading.
Age range	3–5 years
Domains measured	Language/communication skills, general knowledge and awareness
Reliability	Reliability with Family and Child Experiences Survey (FACES) data includes *internal consistency* reflected by Cronbach's alpha of .55 for book knowledge, .71 for print knowledge, and .42 for reading comprehension. *Test-retest reliability* was assessed using a six- to nine-month interim and reflected .41 for book knowledge, .17 for print knowledge, and .29 for reading comprehension.
Validity	For the first cohort of the FACES study (1997–99) validity analyses were conducted for the entire FACES battery. Two outcome variables from the Early Childhood Longitudinal Study–Kindergarten (ECLS–K) were used in these analyses: ECLS–K Reading Scale and ECLS–K General Knowledge Scale. *Predictive validity.* There was a correlation between book knowledge scores at the end of the Head Start year (spring 1998) and ECLS–K Reading Scale scores at the end of the kindergarten year (spring 1999): r = .39. There was also a correlation between book knowledge scores at the end of the Head Start year and ECLS–K General Knowledge Scale scores at end of the kindergarten year: r = .52. In multivariate regression analyses with the scale scores from the entire FACES battery at the end of the Head Start year predicting ECLS–K General Knowledge Scale scores at the end of the kindergarten year, book knowledge was a significant predictor (beta = .06).
Languages available	Translated into Spanish by FACES research team
Administration time	Not available
Training recommended	Trained assessors. Training is required for the standardized administrative procedures of the tasks.
States that used this instrument	Georgia, Missouri, North Carolina

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

25. Teacher-Child Rating Scale (T–CRS)	
Author(s)	Hightower, A. D., Work, W. C., Cowen, E. L., Lotyczewski, B. S., Spinell, A. P., Guare, J. C., & Rohrbeck, C. A.
Publisher	Children's Institute
Web site	www.childrensinstitute.net
Intended purpose	Screening
Age range	Kindergarten through third grade
Domains measured	Social/emotional development
Reliability	*Test-retest reliability* data were gathered on 185 subjects. Reliability coefficients ranged from .61 to .91 on each dimension.
Validity	Validity is based on the scales' ability to discriminate groups known to differ in adjustment, and convergent and divergent validity with other measures. T–CRS consistently differentiated between children adjusting well and poorly. This supports the measure's utility as a screening/assessment and program evaluation tool. Generally, correlations between T–CRS scales and other criterion measures support *construct validity* (Hightower, Work, Cowen, Lotyczewski, Spinell, Guare, & Rohrbeck, 1986).
Languages available	English
Administration time	10 minutes
Training recommended	Users manual gives specific directions. Examiners may call the publisher if they have questions.
States that used this instrument	New York

26. Woodcock-Johnson III (W-J III)	
Author(s)	Woodcock, R., McGrew, K., & Mather, N.
Publisher	Riverside Publishing
Web site	http://www.riverpub.com
Intended purpose	Screening, informing instruction, tracking child outcomes, conducting research
Age range	2–90+ years
Domains measured	Cognitive development, math/pre-math skills, general knowledge and awareness, language/communication skills, literacy skills, overall child development. The specific subtests that were used most often in the evaluations reviewed for this study included Applied Problems, Letter-Word Identification, Math Fluency, and Sound Awareness.
Reliability	In general, the W-J III has reliability coefficients of .80 or higher; several are .90 or higher. *Split-half reliability* coefficients are .93 for Applied Problems, .94 for Letter-Word Identification, .90 for Math Fluency, and .81 for Sound Awareness. Although these are strong reliabilities for individual tests, the interpretive plan is based on cluster interpretation. W-J III clusters show strong reliabilities, most at .90 or higher.
Validity	*Concurrent validity*. In a study of 202 children ages 1 year 9 months to 6 years 3 months data were collected with the W-J III Tests of Cognitive Abilities and Tests of Achievement and with the Wechsler Preschool and Primary Scale of Intelligence–Revised. Correlations with the W-J III and the Wechsler ranged from .53 to .74. A second validity study was conducted with 32 preschool-age children ranging from 3 years to 5 years 10 months. Correlations between scores on the W-J III and the Stanford-Binet (SB) Intelligence Scale–Fourth Edition ranged from .03 to .76, with the lowest correlations with the SB-IV Quantitative Reasoning subscale. The majority of correlations on other subtests were .44 and above, and most were .65 and above. An *internal validity* study examined the extent to which the W-J III tests of similar abilities were more highly correlated with each other than with tests designed to assess different abilities. Using data from the norming sample (with ages 2–3 years and 4–5 years analyzed separately), investigators found that for the most part the expected correlations were evident—subtests designed to test similar abilities were more highly correlated with each other than with subtests measuring abilities associated with a different construct. Results from confirmatory factor analyses were consistent with these results. Using data from children ages six years and older in the norming sample, the investigators found that results generally supported the conceptual model that underlies the subtests.
Languages available	English and Spanish
Administration time	35–45 minutes
Training recommended	Only trained personnel should administer the W-J III.
States that used this instrument	Georgia, Michigan, Missouri, New Jersey, North Carolina, Oklahoma, West Virginia

(CONTINUED)

TABLE D1 (CONTINUED)
Abstracts of key instruments

27. Work Sampling System (WSS)	
Author(s)	Meisels, S., Jablon, J., Dichtelmiller, M., Dorfman, A. & Marsden D.
Publisher	Pearson Early Learning
Web site	www.pearsonearlylearning.com
Intended purpose	Informing instruction, tracking child outcomes
Age range	3 years through grade 6
Domains measured	Cognitive development, math/pre-math skills, general knowledge and awareness, social/emotional development, child health/physical development, language/communication skills
Reliability	Information gathered from the web site cites Meisels, Liaw, Dorfman, & Fails (1995) as evidence of reliability. In this study children in classrooms using the WSS were also given individually administered, norm-referenced assessments in the fall and spring. Results indicated that the Work Sampling Checklist and Summary had high *internal reliability* and moderately high *interrater reliability*.
Validity	Results indicated that the WSS correlates well with the Woodcock-Johnson–Revised. In addition, it is a reliable predictor of achievement ratings in kindergarten through grade 3. The WSS was able to discriminate accurately between children at risk and those who are not (Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 1998).
Languages available	English and Spanish
Administration time	Ongoing curriculum-embedded assessment
Training recommended	Only trained professionals can administer the WSS.
States that used this instrument	Kansas, Maryland, Minnesota, Nebraska

TABLE D2
"At a glance" table for key instruments

Number	Name of instrument	Purpose	Age range	Domains assessed	Type of administration
1.	Basic School Skills Inventory–Third Edition	Aiding teachers in making decisions about programming and instruction	4 years to 6 years 11 months	Basic knowledge, language/communication skills, literacy skills, math/ pre-math skills, classroom behavior	Direct child assessment
2.	Battelle Developmental Inventory	Screening, tracking child outcomes	Infant through 7–11 years	Cognitive development, social/emotional development, language skills, child health/ physical development	Direct child assessment; teacher observation; parent interviews
3.	Bracken Basic Concepts Scale–Revised (BBCS–R)	Assessing academic readiness	2 years 6 months to 7 years 11 months	Cognitive development, language/communication skills, math/pre-math skills, social/emotional development, school readiness	Direct child assessment
4.	California Preschool Social Competency Scale	Assessing children's social adjustment in the classroom	Preschool-age children	Social/emotional development, school adjustment/attitude	Teacher observation
5.	Child Observation Record (COR)	Primarily providing data for teachers to use in planning instruction; also documenting children's progress and evaluating a classroom or program.	2 years 6 months to 6 years	Initiative, social relations, creative representation, movement/music, language/literacy skills, mathematics, science	Teacher observation
6.	Color Bears and Counting Bears	Assessing knowledge of colors and counting ability	3–5 years	Early literacy and numeracy	Direct child assessment
7.	Creative Curriculum Developmental Continuum for Ages 3 – 5 Assessment	Informing instruction, tracking child outcomes, conducting research	3 years through kindergarten	Cognitive development, general knowledge and awareness, social/ emotional development, child health/physical development, language/ communication skills	Teacher observation
8.	Comprehensive Test of Phonological Processing (CTOPP)	Screening, tracking child outcomes, conducting research	5 years through 24 years 11 months	Language/ communication skills	Direct child assessment
9.	Developmental Observation Checklist System	Screening, tracking child outcomes	Birth through first grade	Language/ communication skills, cognitive development, social/emotional development, motor movement	Parent report

(CONTINUED)

TABLE D2 (CONTINUED)
"At a glance" table for key instruments

Number	Name of instrument	Purpose	Age range	Domains assessed	Type of administration
10.	Developing Skills Checklist (DSC)	Planning instruction	4–6 years	Language/ communication skills, visual, auditory, math/ pre-math skills, memory, print and writing, social/ emotional development, fine and gross motor movement	Direct child assessment; teacher observation; parent interview
11.	Developmental Indicators for the Assessment of Learning–Third Edition (DIAL–3)	Screening, informing instruction	3 years though second grade	Social/emotional development, health/ physical development, language/communication skills, self-help, concepts	Direct child assessment
12.	Expressive One-Word Picture Vocabulary Test (EOWPVT)	Screening, monitoring growth, evaluating program effectiveness	2 years through 18 years 11 months	Language/ communication skills	Direct child assessment
13.	Get It! Got It! Go!	Monitoring change	30–66 months	Literacy skills	Direct child assessment
14.	Get Ready to Read	Screening	4-year-olds	Literacy skills	Direct child assessment
15.	Learning Accomplishments Profile–Revised (LAP-R)	Primarily informing instruction, but also tracking progress	36–72 months	Cognitive development, language/communication skills, self-help, motor movement, social skills	Direct child assessment
16.	Oral and Written Language Scales (OWLS)	Screening, informing instruction, conducting research	3–21 years	Cognitive development, language/communication skills, literacy skills	Direct child assessment
17.	Peabody Picture Vocabulary Test–Third Edition (PPVT–III)	Screening, tracking child outcomes, conducting research	2 years 6 months through 17 years 11 months	Cognitive development, language/communication skills	Direct child assessment
18.	Preschool Comprehensive Test of Phonological Processing (Pre-CTOPP)	Assessing children's phonological awareness, phonological memory, and phonological access	3–5 years	Early literacy skills	Direct child assessment
19.	Pre-Language Assessment Scales (Pre-LAS 2000)	Measuring oral language proficiency and pre-literacy skills	4 years through first grade	Cognitive development, language/communication skills, literacy skills	Direct child assessment
20.	Preschool and Kindergarten Behavior Scales (PKBS)	Screening, informing instruction, conducting research	3–6 years	Social/emotional development	Rating scale
21.	Preschool Language Scale–Fourth Edition (PLS–4)	Screening	Infant to 6 years 11 months	Expressive and receptive language	Direct child assessment

Number	Name of instrument	Purpose	Age range	Domains assessed	Type of administration
22.	Receptive One-Word Picture Vocabulary Test (ROWPVT)	Assessing ability to understand spoken and written vocabulary of others	Infant through fourth grade and beyond	Language/ communication skills	Direct child assessment
23.	Social Skills Rating System (SSRS)	Screening, informing instruction, tracking child outcomes, conducting research	3 years through kindergarten	Social/emotional development, academic competence	Rating scale
24.	Story and Print Concepts	Assessing basic story concepts, print concepts, and the mechanics of reading	3–5 years	Language/ communication skills, general knowledge and awareness	Direct child assessment
25.	Teacher–Child Rating Scale (T–CRS)	Screening	Kindergarten through third grade	Social/emotional development	Rating scale
26.	Woodcock-Johnson III	Screening, informing instruction, tracking child outcomes, conducting research	2–90+ years	Cognitive development, math/pre-math skills, general knowledge and awareness, language/ communication skills, literacy skills, overall child development	Direct child assessment
27.	Work Sampling System (WSS)	Informing instruction, tracking child outcomes	3 years through sixth grade	Cognitive development, math/pre-math skills, general knowledge and awareness, social/ emotional development, child health/physical development, language/ communication skills	Observation

REFERENCES

Administration on Children, Youth, and Families. (2003, June 26). Head Start National Reporting System on Child Outcomes (Information Memorandum ACYF-IM-HS-03-07). Retrieved July 24, 2007, from http://eclkc.ohs.acf.hhs.gov/hslc/Program%20Design%20and%20Management/Head%20Start%20Requirements/IMs/2003/resour_ime_00206a_020206.html

Barnett, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *The Future of Children, 5*(3), 25–50.

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

Bridges, L. J., Berry, D. J., Johnson, R., Calkins, J., Margie, N. G., Cochran, S. W., Ling, T. J., & Zaslow, M. J. (2004). *Early childhood measures profiles.* Washington, DC: Child Trends. Retrieved from http://aspe.hhs.gov/HSP/ECMeasures04/index.htm

Brown, E. G., & Scott-Little, C. (2003). *Evaluations of school readiness initiatives: What are we learning?* Tallahassee, FL: SERVE.

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

California Department of Education. (2006, September 13). *Introduction to desired results.* Retrieved February 16, 2007, from http://www.cde.ca.gov/sp/cd/ci/desiredresults.asp

Carrow-Woolfolk, E. (1996). *Oral and Written Language Scales: Written Expression Scale Manual.* Circle Pines, MN: American Guidance Service.

Chapel Hill Training and Outreach Project. (1995). *Learning Accomplishment Profile–Revised.* Chapel Hill, NC: Author.

Cizek, G., & Fairbanks, D. (2004). Developmental Indicators for the Assessment of Learning–Third Edition. *Mental measurements yearbook*. Omaha, NE: University of Nebraska Press, Buros Institute.

Cohen, L. G., & Spenciner, L. J. (2007). *Assessment of children & youth with special needs.* New York: Pearson Allyn and Bacon.

CTB. (1990). *Developing Skills Checklist*. Monterey, CA: McGraw-Hill.

DeMers, S. T., & Fiorello, C. (1999). Legal and ethical issues in preschool assessment and screening. In E. V. Nuttall, I. Romero, & J. Kalesnik (Eds.), *Assessing and screening preschoolers: Psychology and educational dimensions.* Boston: Allyn and Bacon.

Denham, S. A., & Burton, R. (2003). Assessing emotional and social competence during preschool years. In S. A. Denham & R. Burton (Eds.), *Social and emotional prevention and intervention programming for preschoolers.* New York: Kluwer Academic.

Dodge, D., Colker, L., & Heroman, C. (2000). *Connecting content, teaching, and learning.* Washington, DC: Teaching Strategies.

Duncan, S.E., & De Avila, E. (1998). *PreLAS 2000.* Monterey, CA: CTB/McGraw-Hill

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

Early Childhood Research Institute on Measuring Growth and Development. (1998). *Research and development of individual growth and development indicators for children between birth and age eight* (Technical Report No. 4). Minneapolis, MN: Center for Early Education and Development, University of Minnesota.

Epstein, A. S., Schweinhart, L., DeBruin-Parecki, A., & Robin, K. B. (2004). Preschool assessment: A guide to developing a balanced approach. *Preschool Policy Matters*, Issue 7. New Brunswick, NJ: Rutgers University, National Institute for Early Education Research.

Florida Statutes. (2006). § Title XLVIII K – 20 Education Code, 1002 and Title XXX, Social Welfare, 411.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational Research* (6th ed.). White Plains, NY: Longman Publishers.

Gilliam, W. S., & Zigler, E. (2001). A critical meta-analysis of all evaluations of state-funded preschool from 1977 to 1998: Implications for policy, service delivery, and program evaluation. *Early Childhood Research Quarterly, 15*(4), 441–473.

Gilliam, W. S., & Zigler, E. (2004). *State efforts to evaluate the effects of prekindergarten: 1977 to 2003*. New Brunswick, NJ: National Institute for Early Education Research (NIEER). Retrieved May 26, 2006, from http://www.nieer.org/resources/research/StateEfforts.pdf

Graue, E. (2006). The answer is readiness—Now what is the question? *Early Education and Development*, *17*(1), 43–56.

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

Grinder, E. L., & Kochanoff, A. (2006). *Updated report on early childhood assessment for children from birth to age 8 (grade 3).* Harrisburg, PA: Pennsylvania's Departments of Education and Public Welfare. Retrieved from http://www.pde.state.pa.us/early_childhood/lib/early_childhood/Build_9-06_AssessB-8.pdf

Guralnick, M. J. (Ed.). (1997). *The effectiveness of early intervention*. Baltimore, MD: Paul H. Brookes.

Hammill, D., Leigh, J., Pearson, N., & Maddox, T. (1998). *Basic School Skills Inventory–Third Edition.* Austin, TX: Pro-Ed.

Hebbeler, K., Barton, L. R., & Mallik, S. (2007). Assessment and accountability for programs serving young children with disabilities. Chapel Hill, NC: Early Childhood Outcomes Center. Retrieved July 24, 2007, from http://www.fpg.unc.edu/~eco/papers.cfm

Henry, G. T., Henderson, L. W., Ponder, B. P., Gordon, C. S., Mashburn, A., & Rickman, D. K. (2003). *Report of the findings from the Georgia early childhood study: 2001–2002.* Atlanta, GA: Georgia State University.

High/Scope Educational Research Foundation. (1992). *Child Observation Record–Manual*. Ypsilanti, MI: High/Scope Educational Research Foundation.

Hightower, D., Work, W., Cowen, E., Lotyczewski, B., Spinell, A., Guare, J., & Rohrbeck, C. (1986). The Teacher-Child Rating Scale: A brief objective measure of elementary children's school problem behavior and competencies. *School Psychology Review, 15*(3), 393–409.

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

Huffman, L. C., Mehlinger, S. L., & Kerivan, A. S. (2000). *A good beginning: Sending America's children to school with the social and emotional competencies they need to succeed.* Bethesda, MD: Children's Mental Health Foundation and Agencies Network.

Karoly, L. A., Greenwood, P. W., Everingham, S. S., Hoube, J., Kilburn, M. R., Rydell, C. P., Sanders, M., & Chiesa, J. (1998). *Investing in our children: What we know and don't know about the costs and benefits of early childhood interventions*. Washington, DC: RAND Corporation.

Kim, J., & Suen, H. K. (2003). Predicting children's academic achievement from early assessment scores: A validity generalization study. *Early Childhood Research Quarterly, 18*, 547–566.

Kisker, E. E., Boller, K., Nagatoshi, C., Sciarrino, C., Jethwani, V., Zavitsky, T., Ford, M., & Love, J. M. (2003, April). *Resources for measuring services and outcomes in Head Start Programs serving infants and toddlers*. Princeton, NJ: Mathematica Policy Research, Inc. Retrieved from http://www.acf.hhs.gov/programs/opre/ehs/perf_measures/reports/resources_measuring/res_meas_title.html

Lamb, M. E. (1998). Nonparental child care: Context, quality, correlates, and consequences. In I. E. Sigel & K. A. Renninger (Eds.), *Handbook of child psychology.* Vol. 4, *Child psychology in practice.* New York: Wiley.

La Paro, K. M., & Pianta, R. C. (2000). Predicting children's competence in the early school years: A meta-analytic review. *Review of Educational Research*, *70,* 443–484.

Levine, S., Elzey, F. F., & Lewis, M. (1969). *California Preschool Social Competency Scale*. Palo Alto, CA: Consulting Psychologists Press.

Li, C., Walton, J. R., & Nuttall, E. V. (1999). Preschool evaluation of culturally and linguistically diverse children. In E. V. Nuttall, I. Romero, & J. Kalesnik (Eds.), *Assessing and screening preschoolers: Psychological and educational dimensions* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Lonigan, C. J. (2006). Development, assessment, and promotion of preliteracy skills. *Early Education and Development, 17*(1), 91–114.

Maryland Department of Education. (2006). Maryland model for school readiness. Retrieved March 24, 2007, from http://www.mdk12.org/instruction/ensure/MMSR/

McCallion, G. (2004). *Early childhood education: Federal policy issues*. Congressional Research Services: Report for Congress. RL31123. Washington, DC.

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

Meisels, S. J. (1999). Assessing readiness. In R. C. Pianta & M. J. Cox (Eds.), *The transition to kindergarten*. Baltimore, MD: Paul H. Brookes Publishing Co.

Meisels, S., Bickel, D., Nicholson, J., Xue, Y., & 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), 73–95.

Meisels, S., Jablon, J., Dichtelmiller, M., Dorfman, A., & Marsden, D. (1998). *The Work Sampling System*. Ann Arbor, MI: Pearson Early Learning.

Merrell, K. W. (1996). Social-emotional assessment in early childhood: The Preschool and Kindergarten Behavior Scales. *Journal of Early Intervention, 20*(2), 132–145.

National Education Goals Panel. (1995). *Reconsidering children's early development and learning: Toward common views and vocabulary.* Goal 1 Technical Planning Group. S. L. Kagan, E. Moore, & S. Bredekamp (Eds.). Washington, DC: U.S. Government Printing Office.

National Education Goals Panel. (1998). *Principles and recommendations for early childhood assessments*. Washington, DC: Author.

National Research Council. (2001). *Eager to learn: Educating our preschoolers.* Committee on Early Childhood Pedagogy. B. T. Bowman, M. S. Donovan, & M. S. Burns (Eds.). Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council and Institute of Medicine. (2000). *From neurons to neighborhoods: The science of early childhood development*. Committee on Integrating the Science of Early Childhood Development. In J. P. Shonkoff & D. A. Phillips (Eds.), *Board of Children, Youth, and Families, Commission on Behavioral and Social Sciences and Education*. Washington, DC: National Academy Press.

Newborg, J., Stock J., Wnek, L., Guidubaldi, J., & Svinicki, J. (1984). *Battelle Developmental Inventory with recalibrated technical data and norms: examiner's manual.* Rolling Meadows, IL: Riverside Publishing.

Pearson Assessments. (2007). Test de Vocabulario en Imagenes Peabody. Retrieved February 8, 2007, from http://ags.pearsonassessments.com/group.asp?nGroupInfoID=a2600.

Ponder, B. D., Rickman, D. K., & Henry, G. T. (2004). *Evaluation of the pre-K summer readiness pilot program.* Atlanta, GA: Georgia State University.

Pre[k]now. (2006). *Presentations: Pre-K across the nation.* Retrieved November 23, 2006, from www.preknow.org/documents/presentations/Pre-K_Across_the_USA.pdf

Sattler, J. M. (2002). *Assessment of children* (3rd ed.). La Mesa, CA: Jerome Sattler.

Schweinhart, L. (2003). Issues in implementing a state preschool program evaluation. In C. Scott-Little, S. L. Kagan, & R. Clifford (Eds.), *Assessing the state of state assessments: Perspectives on assessing young children.* Tallahassee, FL: SERVE.

Scott-Little, C., & Niemeyer, J. (2001). *Assessing kindergarten children: What school systems need to know.* Tallahassee, FL: SERVE.

Snow, K. L. (2006). Measuring school readiness: Conceptual and practical considerations. *Early Education and Development*, *17*(1), 7– 41.

Spies, R. A., & Plake, B. S. (2005). *The sixteenth mental measurements yearbook* [Electronic version]. Lincoln, NE: Buros Institute of Mental Measurements. Retrieved November 18, 2006, from http://www.unl.edu/buros

Trochim, W. M. (2006). The research methods knowledge base (2nd ed.). Retrieved February 14, 2007, from http://www.socialresearchmethods.net/kb/reltypes.php

Wagner, R., Torgeson, J., & Rashotte, C. (1999). *Comprehensive Test of Phonological Processing (CTOPP).* Austin, TX: Pro-Ed.

Whitehurst, G. (2006). The NCLD Get Ready to Read Screening Tool technical report. Retrieved July 24, 2007, from http://www.getreadytoread.org/index.php?option=content&task=view&id=80&Itemid=53.

Zaslow, M., & Halle, T. (2003). Statewide school readiness assessments: Challenges and next steps. In C. Scott-Little, S. L. Kagan, & R. Clifford (Eds.), *Assessing the state of state assessments: Perspectives on assessing young children.* Tallahassee, FL: SERVE.

Zaslow, M., Halle, T., Martin, L., Cabrera, N., Calkins, J., Pitzer, L., & Margie, N. G. (2006). Child outcome measures in the study of child care quality. *Education Review, 30*, 577–610.

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