PRINCIPLES AND RECOMMENDATIONS FOR EARLY CHILDHOOD ASSESSMENTS

Submitted to

THE NATIONAL EDUCATION GOALS PANEL

by the Goal 1 Early Childhood Assessments Resource Group

Lorrie Shepard, Sharon Lynn Kagan, and Emily Wurtz, Editors

National Education Goals Panel

Governors

James B. Hunt, Jr., North Carolina (Chair, 1997–1998)
John Engler, Michigan
William Graves, Kansas
Paul E. Patton, Kentucky
Roy Romer, Colorado
Tommy G. Thompson, Wisconsin
Cecil Underwood, West Virginia
Christine Todd Whitman, New Jersey

Members of the Administration

Carol H. Rasco, Senior Advisor to the Secretary of Education
Richard W. Riley, Secretary of Education

Members of Congress

U.S. Senator Jeff Bingaman, New Mexico
U.S. Senator Jim Jeffords, Vermont
U.S. Representative William F. Goodling, Pennsylvania
U.S. Representative Dale E. Kildee, Michigan

State Legislators

Representative G. Spencer Coggs, Wisconsin
Representative Ronald Cowell, Pennsylvania
Representative Mary Lou Cowlishaw, Illinois
Representative Douglas R. Jones, Idaho

National Education Goals Panel Staff

Ken Nelson, Executive Director
Leslie A. Lawrence, Senior Education Associate
Cynthia D. Prince, Associate Director for Analysis and Reporting
Emily O. Wurtz, Senior Education Associate
Cynthia M. Dixon, Program Assistant
John Masaitis, Executive Officer
Sherry Price, Secretary

Goal 1 Early Childhood Assessments Resource Group

Leaders: Sharon Lynn Kagan, Yale University
Lorrie Shepard, University of Colorado

Sue Bredekamp, National Association for the Education of Young Children
Edward Chittenden, Educational Testing Service
Harriet Egertson, Nebraska State Department of Education
Eugene García, University of California, Berkeley
M. Elizabeth Graue, University of Wisconsin
Kenji Hakuta, Stanford University
Carollee Howes, University of California, Los Angeles
Annemarie Palincsar, University of Michigan
Tej Pandey, California State Department of Education
Catherine Snow, Harvard University
Maurice Sykes, District of Columbia Public Schools
Valora Washington, The Kellogg Foundation
Nicholas Zill, Westat, Inc.

February 1998

PRINCIPLES AND RECOMMENDATIONS FOR EARLY CHILDHOOD ASSESSMENTS

Goal 1: Ready to Learn

By the year 2000, all children in America will start school ready to learn.

Objectives:

- All children will have access to high-quality and developmentally appropriate preschool programs that help prepare children for school.
- Every parent in the United States will be a child's first teacher and devote time each day to helping such parent's preschool child learn, and parents will have access to the training and support parents need.
- Children will receive the nutrition, physical activity experiences, and health care needed to arrive at school with healthy minds and bodies, and to maintain the mental alertness necessary to be prepared to learn, and the number of low-birthweight babies will be significantly reduced through enhanced prenatal health systems.

TITLE II—NATIONAL EDUCATION REFORM LEADERSHIP, STANDARDS, AND ASSESSMENTS

PART A—NATIONAL EDUCATION GOALS PANEL

SEC. 201. PURPOSE.

It is the purpose of this part to establish a bipartisan mechanism for—

SEC. 207. EARLY CHILDHOOD ASSESSMENT.

(a) IN GENERAL.—The Goals Panel shall support the work of its Resource and Technical Planning Groups on School Readiness (hereafter in this section referred to as the "Groups") to improve the methods of assessing the readiness of children for school that would lead to alternatives to currently used early childhood assessments.

(b) ACTIVITIES.—The Groups shall—

(1) develop a model of elements of school readiness that address a broad range of early childhood developmental needs, including the needs of children with disabilities;

(2) create clear guidelines regarding the nature, functions, and uses of early childhood assessments, including assessment formats that are appropriate for use in culturally and linguistically diverse communities, based on model elements of school readiness;

(3) monitor and evaluate early childhood assessments, including the ability of existing assessments to provide valid information on the readiness of children for school; and

(4) monitor and report on the long-term collection of data on the status of young children to improve policy and practice, including the need for new sources of data necessary to assess the broad range of early childhood developmental needs.

(c) ADVICE.—The Groups shall advise and assist the Congress, the Secretary, the Goals Panel, and others regarding how to improve the assessment of young children and how such assessments can improve services to children.

(d) REPORT.—The Goals Panel shall provide reports on the work of the Groups to the appropriate committees of the Congress, the Secretary, and the public.

Introduction

Americans want and need good information on the well-being of young children. Parents want to know if their children will be ready for school. Teachers and school administrators want to know if their programs are effective and if they are providing children the right programs and services. Policymakers want to know which program policies and expenditures will help children and their families, and whether they are effective over time. Yet young children are notoriously difficult to assess accurately, and well-intended testing efforts in the past have done unintended harm. The principles and recommendations in this report were developed by advisors to the National Education Goals Panel to help early childhood professionals and policymakers meet their information needs by assessing young children appropriately and effectively.

The first National Education Goal set by President Bush and the nation's Governors in 1990 was that by the year 2000, all children in America will start school ready to learn. This Goal was meant to help those advocating the importance of children's needs. Yet from the start, Goal 1 proved problematic to measure. The Panel could find no good data or methods to measure children's status when they started school. In view of the importance of this issue, Congress in 1994 charged the Goals Panel to support its Goal l advisors to "*create clear guidelines regarding the nature, functions, and uses of early childhood assessments, including assessment formats that are appropriate for use in culturally and linguistically diverse communities, based on model elements of school readiness.*" The principles and recommendations in this document are the result of efforts by the Goal 1 Early Childhood Assessments Resource Group to address this charge.

Assessment and the Unique Development of Young Children

Assessing children in the earliest years of life—from birth to age 8—is difficult because it is the period when young children's rates of physical, motor, and linguistic development outpace growth rates at all other stages. Growth is rapid, episodic, and highly influenced by environmental supports: nurturing parents, quality caregiving, and the learning setting.

Because young children learn in ways and at rates different from older children and adults, we must tailor our assessments accordingly. Because young children come to know things through doing as well as through listening, and because they often represent their knowledge better by showing than by talking or writing, paper-and-pencil tests are not adequate. Because young children do not have the experience to understand what the goals of formal testing are, testing interactions may be very difficult or impossible to structure appropriately. Because young children develop and learn so fast, tests given at one point in time may not give a complete picture of learning. And because young children's achievements at any point are the result of a complex mix of their ability to learn and past learning opportunities, it is a mistake to interpret measures of past learning as evidence of what could be learned.

For these reasons, how we assess young children and the principles that frame such assessments need special attention. What works for older children or adults will not work for younger children; they have unique needs that we, as adults, are obliged to recognize if we are to optimize their development.

Recent Assessment Issues

Educators and child development specialists have long recognized the uniqueness of the early years. Informal assessment has characterized the early childhood field. Early educators have observed and recorded children's behavior naturalistically, watching children in their natural environments as youngsters carry out everyday activities. These observations have proven effective for purposes of chronicling children's development, cataloging their accomplishments, and tailoring programs and activities within the classroom to meet young children's rapidly changing needs.

Recently, however, there has been an increase in formal assessments and testing, the results of which are used to make "high stakes" decisions such as tracking youngsters into high- and low-ability groups, (mis)labeling or retaining them, or using test results to sort children into or out of kindergarten and preschools. In many cases, the instruments developed for one purpose or even one age group of children have been misapplied to other groups. As a result, schools have often identified as "not yet ready" for kindergarten, or as "too immature" for group settings, large proportions of youngsters (often boys and non-English speakers) who would benefit enormously from the learning opportunities provided in those settings. In particular, because the alternative treatment is often inadequate, screening out has fostered inequities, widening—and perpetuating—the gap between youngsters deemed ready and unready.

The Current Climate

Despite these difficulties, demands for assessments of student learning are increasing. Pressed by demands for greater accountability and enhanced educational performance, states are developing standards for school-aged children and are creating new criteria and approaches for assessing the achievement of challenging academic goals. In this context, calls to assess young children—from birth through the earliest grades in school—are also increasing. This document attempts to indicate how best to craft such assessments in light of young children's unique development, recent abuses of testing, and the legitimate demands from parents and the public for clear and useful information.

The principles and recommendations in this document are meant to help state and local officials meet their information needs well. They indicate both general principles and specific purposes for assessments, as well as the kinds of provisions needed to ensure that the results will be accurate and useful for those purposes. Because testing young children has in the past led to unfair or harmful effects, the recommendations include warnings to protect against potential misuse. To explain the basis of these recommendations, there is a definition of each of four categories of assessment purpose, the audiences most concerned with the results of each, the technical requirements that each assessment must meet, and how assessment considerations for each purpose vary across the age continuum from birth to 8 years of age.

General Principles

The following general principles should guide both policies and practices for the assessment of young children.

- **Assessment should bring about benefits for children.**
  Gathering accurate information from young children is difficult and potentially stressful. Formal assessments may also be costly and take resources that could otherwise be spent directly on programs and services for young children. To warrant conducting assessments, there must be a clear benefit—either in direct services to the child or in improved quality of educational programs.
- **Assessments should be tailored to a specific purpose and should be reliable, valid, and fair for that purpose.**
  Assessments designed for one purpose are not necessarily valid if used for other purposes. In the past, many of the abuses of testing with young children have occurred because of misuse. The recommendations in the sections that follow are tailored to specific assessment purposes.
- **Assessment policies should be designed recognizing that reliability and validity of assessments increase with children's age.**
  The younger the child, the more difficult it is to obtain reliable and valid assessment data. It is particularly difficult to assess children's cognitive abilities accurately before age 6. Because of problems with reliability and validity, some types of assessment should be postponed until children are older, while other types of assessment can be pursued, but only with necessary safeguards.

- **Assessments should be age-appropriate in both content and the method of data collection.**
  Assessments of young children should address the full range of early learning and development, including physical well-being and motor development; social and emotional development; approaches toward learning; language development; and cognition and general knowledge. Methods of assessment should recognize that children need familiar contexts in order to be able to demonstrate their abilities. Abstract paper-and-pencil tasks may make it especially difficult for young children to show what they know.
- **Assessments should be linguistically appropriate, recognizing that to some extent all assessments are measures of language.**
  Regardless of whether an assessment is intended to measure early reading skills, knowledge of color names, or learning potential, assessment results are easily confounded by language proficiency, especially for children who come from home backgrounds with limited exposure to English, for whom the assessment would essentially be an assessment of their English proficiency. Each child's first- and second-language development should be taken into account when determining appropriate assessment methods and in interpreting the meaning of assessment results.
- **Parents should be a valued source of assessment information, as well as an audience for assessment results.**
  Because of the fallibility of direct measures of young children, assessments should include multiple sources of evidence, especially reports from parents and teachers. Assessment results should be shared with parents as part of an ongoing process that involves parents in their child's education.

Important Purposes of Assessment for Young Children

The intended use of an assessment—its purpose—determines every other aspect of how the assessment is conducted. Purpose determines the content of the assessment (What should be measured?); methods of data collection (Should the procedures be standardized? Can data come from the child, the parent, or the teacher?); technical requirements of the assessment (What level of reliability and validity must be established?); and, finally, the stakes or consequences of the assessment, which in turn determine the kinds of safeguards necessary to protect against potential harm from fallible assessment-based decisions.

For example, if data from a statewide assessment are going to be used for school accountability, then it is important that data be collected in a standardized way to ensure comparability of school results. If children in some schools are given practice ahead of time so that they will be familiar with the task formats, then children in all schools should be provided with the same practice; teachers should not give help during the assessment or restate the questions unless it is part of the standard administration to do so; and all of the assessments should be administered in approximately the same week of the school year. In contrast, when a teacher is working with an individual child in a classroom trying to help that child learn,

assessments almost always occur in the context of activities and tasks that are already familiar, so practice or task familiarity is not at issue. In the classroom context, teachers may well provide help while assessing to take advantage of the learning opportunity and to figure out exactly how a child is thinking by seeing what kind of help makes it possible to take the next steps. For teaching and learning purposes, the timing of assessments makes the most sense if they occur on an ongoing basis as particular skills and content are being learned. Good classroom assessment is disciplined, not haphazard, and, with training, teachers' expectations can reflect common standards. Nonetheless, assessments devised by teachers as part of the learning process lack the uniformity and the standardization that is necessary to ensure comparability, essential for accountability purposes.

Similarly, the technical standards for reliability and validity are much more stringent for high-stakes accountability assessment than for informal assessments used by individual caregivers and teachers to help children learn. The consequences of accountability assessments are much greater, so the instruments used must be sufficiently accurate to ensure that important decisions about a child are not made as the result of measurement error. In addition, accountability assessments are usually "one-shot," stand-alone events. In contrast, caregivers and teachers are constantly collecting information over long periods of time and do not make high-stakes decisions. If they are wrong one day about what a child knows or is able to do, then the error is easily remedied the next day.

Serious misuses of testing with young children occur when assessments intended for one purpose are used inappropriately for other purposes. For example, the content of IQ measures intended to identify children for special education is not appropriate content to use in planning instruction. At the same time, assessments designed for instructional planning may not have sufficient validity and technical accuracy to support high-stakes decisions such as placing children in a special kindergarten designated for at-risk children.

An appropriate assessment *system* may include different assessments for different categories of purpose, such as:

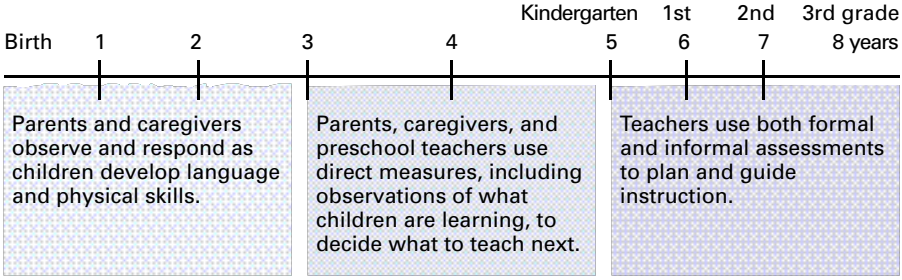
- assessments to support learning,
- assessments for identification of special needs,
- assessments for program evaluation and monitoring trends, and
- assessments for high-stakes accountability.

In the sections that follow, the requirements for each of these assessment purposes are described. Only under special circumstances would it be possible to serve more than one purpose with the same assessment, and then usually at greater cost, because the technical requirements of each separate purpose must still be satisfied. We address the issue of combining purposes in the last section.

Photo: Martin Deutsch

Samples of student work illustrating progress on an emergent writing continuum
(from the North Carolina Grades 1 and 2 Assessment)

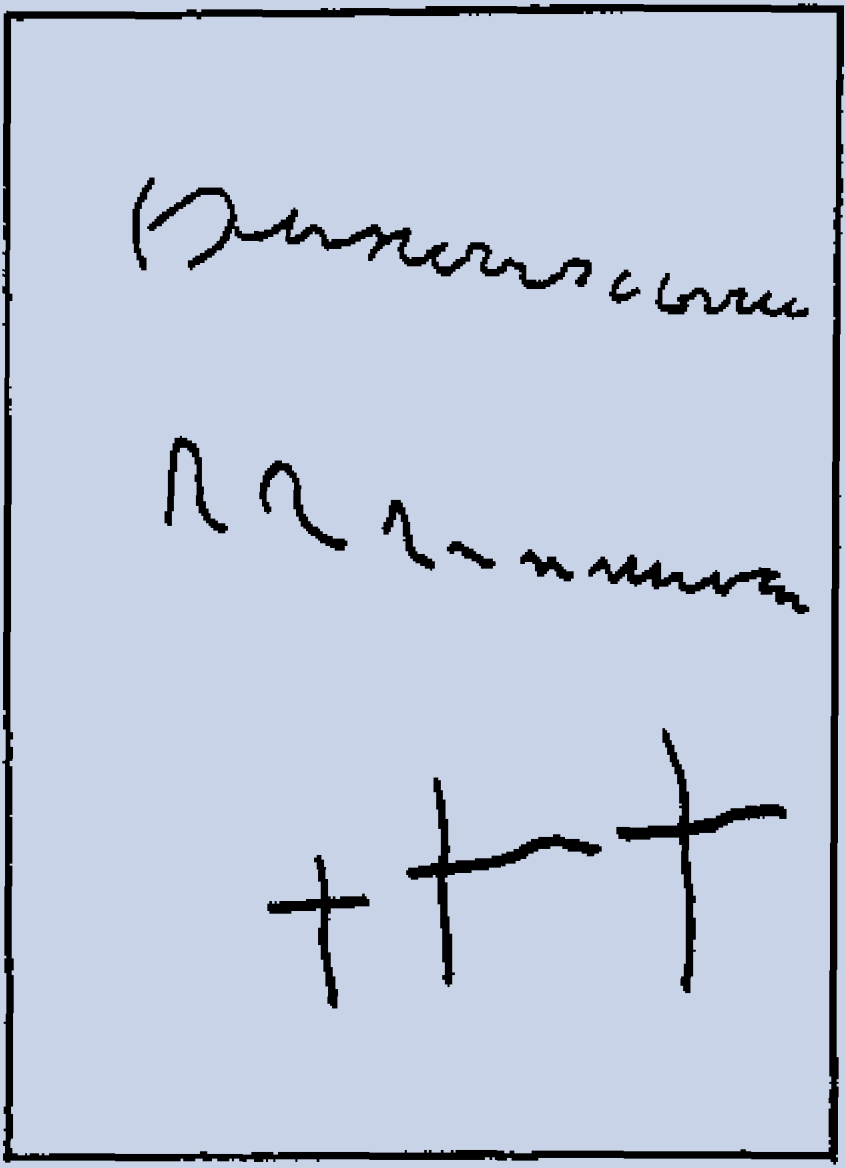
Purpose 1. Assessing to promote children's learning and development



Definition of purpose. Assessing and teaching are inseparable processes. When children are assessed as part of the teaching-learning process, then assessment information tells caregivers and teachers what each child can do and what he or she is ready to learn next. For example, parents watch an infant grow stronger and more confident in walking while holding on to furniture or adults. They "assess" their child's readiness to walk and begin to encourage independent walking by offering outstretched hands across small spaces. In the same vein, preschool teachers and primary-grade teachers use formal and informal assessments to gauge what things children already know and understand, what things could be understood with more practice and experience, and what things are too difficult without further groundwork. This may include appropriate use of early learning readiness measures to be used in planning next steps in instruction. Teachers also use their assessments of children's learning to reflect on their own teaching practices, so that they can adjust and modify curricula, instructional activities, and classroom routines that are ineffective.

Audience. The primary audience for assessments used to support learning is *the teacher,* recognizing, of course, that parents are each child's first teachers. The primary caregiver is asking himself questions about what the child understands, what she does not understand, what she should be learning, and what is too soon for her to be learning, so that the caregiver is constantly providing children with opportunities to learn that are closely congruent with where they are on a learning continuum. In more structured settings, classroom assessments are used by teachers on a ongoing basis to plan and guide instruction. Teachers use both formal and informal assessment information to figure out what is working and to identify which children need additional help.

Children and parents are also important audiences for assessment data gathered as part of instruction. Children benefit from seeing samples of their own work collected over time and from being able to see their own growth and progress. Once children are in the primary grades, helping them become good self-assessors is a valuable skill that helps in future learning. For example, more and more



Sample of student work: the North Carolina Grades 1 and 2 Assessment

teachers are now actively involving children in sharing their accomplishments with parents during conferences. Parents also want and need good information about how their child is doing. Although teachers collect much more information than can be shared with parents, samples of student work and teacher appraisals of each child's progress should be shared on a regular basis as part of an ongoing, reciprocal relationship between professionals and parents. Documentation of children's work with accompanying evaluations helps parents learn about the curriculum and appropriate expectations, as well as their own child's performance. Exchange of information can also be the occasion for parents to offer observations on similar or dissimilar behaviors and skills displayed in home and school contexts.

Principals and primary-grade teachers may also work together to review instructional assessments to make sure that the school's programs are succeeding in helping young children meet developmental and academic expectations. Although external accountability testing should be postponed until third grade because of the difficulties in testing young children, grade-level teams of teachers and school administrators can use instructional assessments for purposes of internal, professional accountability to make sure that children who are struggling receive special help, to identify needs for further professional training, and to improve curricula and instruction.

Policymakers at the state and district level are not the audience for the results of classroom-level assessments. However, policymakers have a legitimate interest in knowing that such assessments are being used at the school level to monitor student learning and to provide targeted learning help, especially for children who are experiencing learning difficulties, such as learning to read. While external accountability testing is not appropriate for first and second graders, policymakers may wish to require that schools have plans in place to monitor student progress and to identify and serve children in need of more intensified help.

Technical requirements. In order for assessments to support learning and development, the content of classroom assessments must be closely aligned with what children are learning, and the timing of assessments must correspond to the specific days and weeks when children are learning particular concepts. Often, this means that informal assessments are made by observing children during an instructional activity. To use assessment information effectively, caregivers and teachers must have enough knowledge about child development and cultural variations to be able to understand the meaning of a child's response and to locate it on a developmental continuum. One example of how children's writing typically develops from scribbles to letters to partially formed words to complete sentences is shown on page 8. Teachers must know not only the typical progression of children's growing proficiency, but also must be sufficiently familiar with age and grade expectations to know when partially formed words would be evidence of precocious performance and when they would be evidence of below-expectation performance that requires special attention and intervention. More formal

assessments, conducted to improve learning, must also be tied to the preschool or primary curriculum and should have clear implications for what to do next.

The reliability and validity requirements for assessments used to support learning are the least stringent of any of the assessment purposes. Over time, teachers' assessments become reliable and consequential, in the sense that multiple assessment events and occasions yield evidence of patterns or consistencies in a child's work, but the day-to-day decisions that caregivers and teachers make on the basis of single assessments are low-stakes decisions. If an incorrect decision is made, for example in judging a child's reading level to help select a book from the library (this book is too easy), that decision is easily changed the next day when new assessment data are available. Because assessments used as part of learning do not have to meet strict standards for technical accuracy, they cannot be used for external purposes, such as school accountability, unless they are significantly restructured. They may, however, inform a school faculty of the effectiveness of its primary school program.

Age continuum. How old a child is within the early childhood age span of birth to 8 years old affects both the what and how of assessment. At all ages, attention should be paid to all five of the dimensions of early learning and development identified by the Goals Panel's Goal 1 Technical Planning Group: physical well-being and motor development; social and emotional development; approaches toward learning; language development; and cognition and general knowledge. Parents of toddlers and early caregivers address all five areas. Beginning in first grade, greater emphasis is placed on language and literacy development and other cognitive-academic domains, though assessment in other domains may continue. Ideally, there should not be an abrupt shift in assessment and instruction from kindergarten to first grade. Instead, preschool assessments used as part of teaching should introduce age-appropriate aspects of emergent literacy and numeracy curricula; and in Grades 1 to 3, physical, social-emotional, and disposition learning goals should continue to be part of classroom teaching and observation.

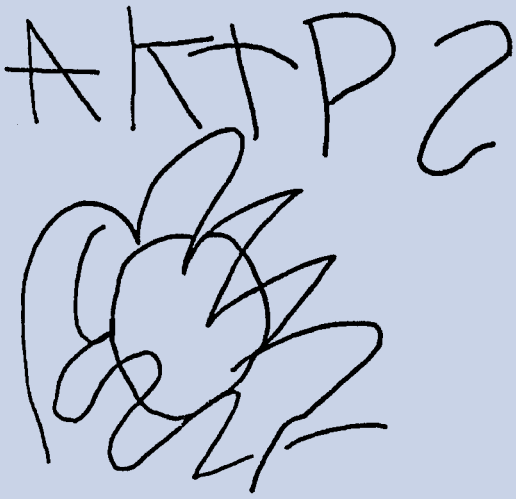
Methods of collecting assessment data include direct observation of children during natural activities; looking at drawings and samples of work; asking questions either orally or in writing; or asking informed adults about the child. The younger the child, the more appropriate it is to use observation. As age increases, especially by third grade, the frequency of more formal assessment "events" should increase, but should still be balanced with informal methods. Across this early childhood age span, children should be introduced to and become comfortable with the idea that adults ask questions and check on understanding as a natural part of the learning process.

Recommendations for what policymakers can do

1. Policymakers should develop or identify assessment materials, to be used instructionally, that exemplify important and age-appropriate learning goals. At the earliest ages, caregivers need tools to assist in observing children. Lacking such assessment materials, preschool programs may misuse screening measures for such purposes. Many local schools and districts lack the resources to develop curricula and closely aligned assessments consistent with standards-based reforms and new Title I requirements. In order for assessment results to be useful instructionally, they should be tied to clear developmental or knowledge continua, with benchmarks along the way to illustrate what progress looks like. Because it is too great an undertaking for individual teachers or early childhood programs to develop such materials on their own, efforts coordinated at the state level can make a significant improvement in assessment practices.

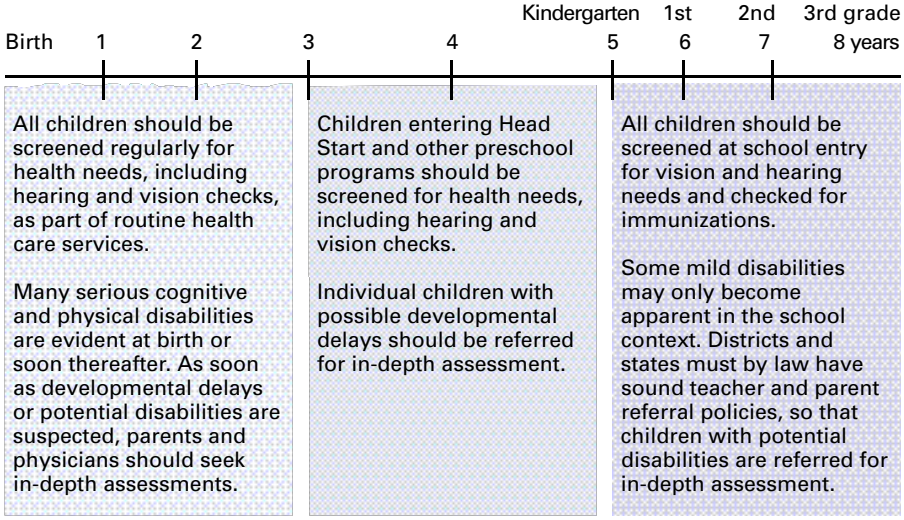
2. Policymakers should support professional development. Early childhood care providers and teachers need better training in children's development within curricular areas in order to be effective in supporting children's learning. Deep understanding of subject matter enables teachers to capitalize on naturally occurring opportunities to talk about ideas and extend children's thinking. In order to make sense of what they are observing, caregivers and teachers need a clear understanding of what typical development looks like in each of the five dimensions, and they also need to understand and appreciate normal variation. When is a child's departure from an expected benchmark consistent with linguistic or cultural differences, and when is it a sign of a potential learning disorder? Teachers and caregivers also need explicit training in how to use new forms of assessment—not only to judge a child's progress, but to evaluate and improve their own teaching practices. Many times, teachers collect children's work in portfolios, but do not know how to evaluate work against common criteria. Or teachers may know how to mark children's papers for right and wrong answers, but need additional training to learn how to document children's thinking, to understand and analyze errors in thinking, and to build on each child's strengths.

Photo: Marietta Lynch



Sample of student work: the North Carolina Grades 1 and 2 Assessment

Purpose 2. Identifying children for health and special services



Definition of purpose. Assessments described in Purpose 1 are used by caregivers and teachers as part of supporting normal learning and development. Assessments used for Purpose 2 help to identify special problems and to determine the need for additional services beyond what regular caregivers can provide. The purpose of identification is to secure special services. Purpose 2 refers to identification of disabilities such as blindness, deafness, physical disabilities, speech and language impairment, serious emotional disturbance, mental retardation, and specific learning disabilities. It also refers to more routine checks for vision, hearing, and immunization to ensure that appropriate health services are provided.

Because of the potential inaccuracy of nearly all sensory and cognitive measures and the cost of in-depth assessments, identification of special needs usually occurs in two stages. *Screening* is the first step in the identification process. It involves a brief assessment to determine whether referral for more in-depth assessment is needed. Depending on the nature of the potential problem, the child is then referred to a physician or child-study team for a more complete evaluation. For mental retardation and other cognitive disabilities, the second-stage in-depth assessment is referred to as a *developmental assessment*.

Audience. The audience for the results of special-needs assessments are the adults who work most closely with the child: the specialists who conducted the assessment and who must plan the follow-up treatment and intervention; parents who must be involved in understanding and meeting their child's needs; and the preschool or primary-grade teacher who works with the child daily and who, most likely, made the referral seeking extra help.

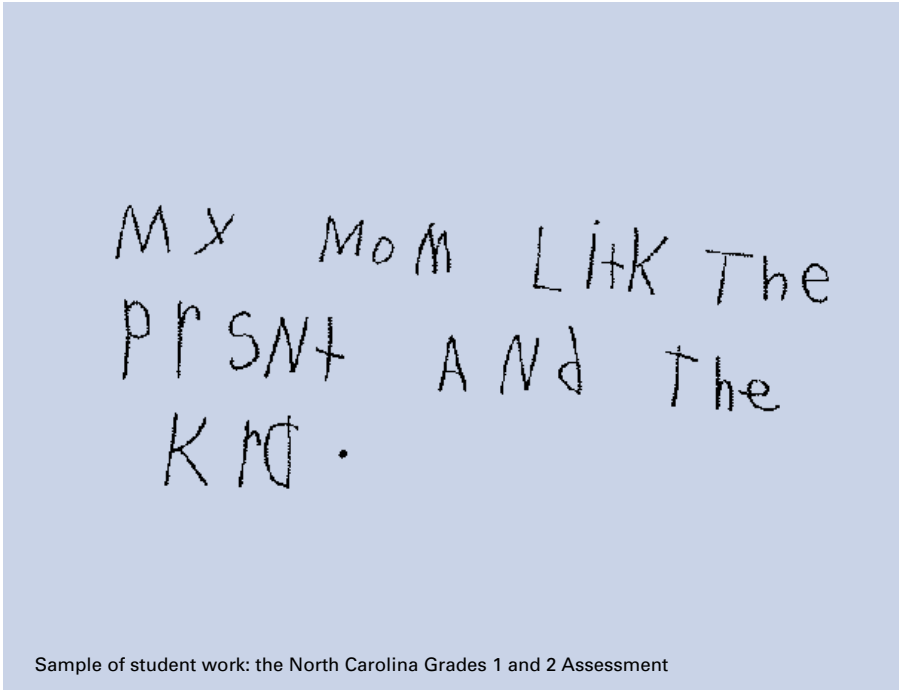
Technical requirements. Except for extreme disabilities, accurate assessment of possible sensory or cognitive problems in young children is very difficult. The instruments used are fallible, and children themselves vary tremendously in their responses from one day to the next or in different contexts. In the field of special education, there is a constant tension between the need to identify children with disabilities to ensure early intervention and help, versus the possible harm of labeling children and possibly assigning them to ineffective treatments.

At step one in the identification process, the screening step, there are two general sources of inaccuracy. First, the instruments are, by design, quick, shortened versions of more in-depth assessments, and are therefore less reliable. Second, they are not typically administered by specialists qualified to make diagnostic decisions. The two-step process is cost-effective, practical, and makes sense so long as the results are only used to signal the need for follow-up assessment. The following warnings are highlighted to guard against typical misuses of screening instruments:

- Screening measures are only intended for the referral stage of identification. They are limited assessments, and typically are administered by school personnel who are not trained to make interpretations about disabilities.

- Screening measures should never be the sole measure used to identify children for special education. Because screening instruments have content like IQ tests, they should also not be used for instructional planning.

For physical disabilities such as vision or hearing impairment, the second-stage in-depth assessment involves more sophisticated diagnostic equipment and the clinical skills of trained specialists. For potential cognitive and language disabilities, the second stage of identification involves trained specialists and more extensive data collection, but, even so, diagnostic procedures are prone to error. To protect against misidentification in either direction (excluding children with disabilities from services or mislabeling children as disabled who are not), several safeguards are built into the identification process for cognitive and language disorders: (1) the sensory, behavioral, and cognitive measures used as part of the in-depth assessment must meet the highest standards of reliability and validity; (2) assessments must be administered and interpreted by trained professionals; (3) multiple sources of evidence must be used and should especially represent competence in both home and school settings; and (4) for children with more than one language, primary language assessments should be used to ensure that language difference is not mistaken for disability. As noted in the age continuum section that follows, screening and identification efforts should be targeted for appropriate ages, taking into account the accuracy of assessment by age.



Sample of student work: the North Carolina Grades 1 and 2 Assessment

Age continuum. Special needs identification starts with the most severe—and most easily recognizable—problems and then identifies children with milder problems as they get older. Children with severe disabilities such as deafness, blindness, severe mental retardation, and multiple physical disabilities are usually identified within the first year of life by parents and physicians. Universal screening of all infants is not recommended, because sensory and cognitive assessments are inaccurate at too early an age, but every child should have access to a regular health care provider, and children should be promptly referred if parents and physicians see that they are not reaching normal developmental milestones.

A referral mechanism contributes to the accuracy of follow-up assessments by serving as an additional data source and checkpoint. As children enter preschool, individual children with possible developmental delays should be referred for in-depth assessment. Some mild disabilities may only become apparent in the school context or, in fact, may only be a problem because of the demands of the school setting. Again, indications of problems should lead to referral for in-depth assessment. Universal hearing and vision screening programs are usually targeted for kindergarten or first grade to ensure contact with every child. Such programs are intended to check for milder problems and disabilities that have gone undetected. For example, if a child has not received regular health checkups, a routine kindergarten screening may uncover a need for glasses.

Sample of student work: the North Carolina Grades 1 and 2 Assessment

Recommendations for what policymakers can do

1. States should ensure that all children have access to a regular health care provider to check for developmental milestones and to ensure that children are on schedule for immunizations by age 2. In addition, states should provide vision and hearing screenings for all children by age 6.

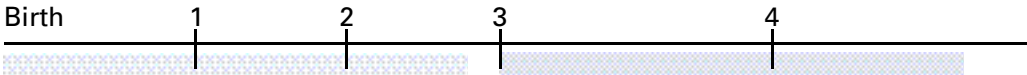
2. The Individuals with Disabilities Education Act requires states to have Child Find programs in place and adequate referral mechanisms in both preschool and the primary grades to ensure that children with potential disabilities are referred for in-depth assessments. Child Find is typically an organized effort by public health, social welfare, and educational agencies to identify all disabled children in need of services.

3. Mild forms of cognitive and language disabilities are particularly hard to identify. We know, however, that effective treatments for children with mild cognitive and language disabilities and most children at-risk for significant reading difficulty all involve the same kinds of high quality, intensive language and literacy interventions. Therefore, policymakers should consider increasing the availability and intensity of such services for broader populations of students who are educationally at-risk, including children in poverty and children thought to have special learning needs.

4. Given the potential for misuse of screening measures, states and districts that mandate screening tests should consider how they are being used and should evaluate whether identifications in their jurisdiction are more accurate with the use of formal tests than in states or districts where only parent and teacher referrals are used.

5. States that mandate administration of cognitive screening measures should expressly forbid the use of screening tests for other than referral purposes. Specifically, screening tests should not be used as readiness tests to exclude children from school; they should not be used to track children by ability in kindergarten and first grade; and they should not be used to plan instruction unless a valid relationship with local curricula has been established.

Appropriate Uses and Technical Accuracy of Assessments Change Across



Purpose 1: Assessing to promote children's learning and development

Parents and caregivers observe and respond as children develop language and physical skills.

Parents, caregivers, and preschool teachers use direct measures, including observations of what children are learning, to decide what to teach next.

Purpose 2: Identifying children for health and special services

All children should be screened regularly for health needs, including hearing and vision checks, as part of routine health care services.

Many serious cognitive and physical disabilities are evident at birth or soon thereafter. As soon as developmental delays or potential disabilities are suspected, parents and physicians should seek in-depth assessments.

Children entering Head Start and other preschool programs should be screened for health needs, including hearing and vision checks.

Individual children with possible developmental delays should be referred for in-depth assessment.

Purpose 3: Monitoring trends and evaluating programs and services

Because direct measures of children's language and cognitive functioning are difficult to aggregate accurately for ages from birth to 2, state reporting systems should focus on living and social conditions that affect learning and the adequacy of services.

Assessments, including direct and indirect measures of children's physical, social, emotional, and cognitive development, could be constructed and used to evaluate prekindergarten programs, but such measures would not be accurate enough to make high-stakes decisions about individual children.

Purpose 4: Assessing academic achievement to hold individual students,

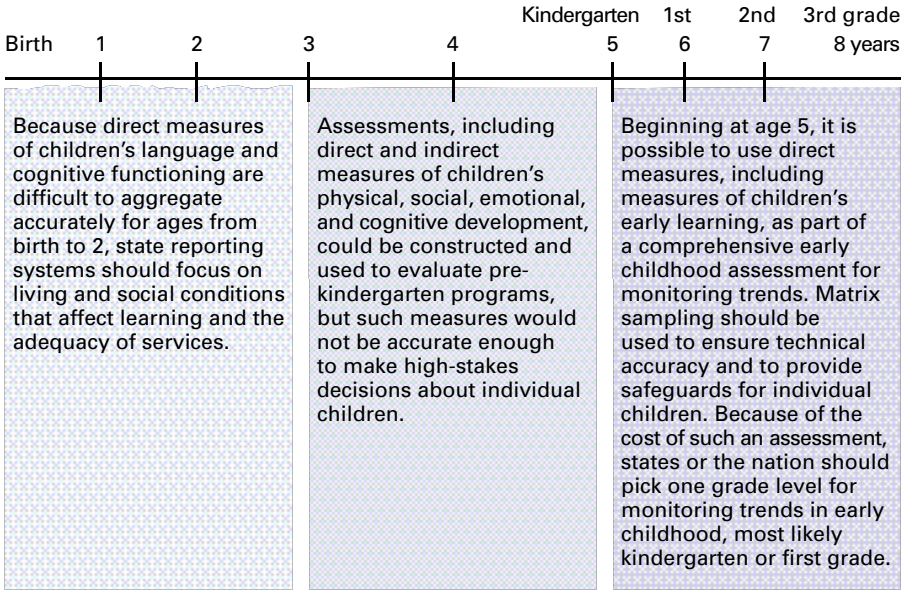
the Early Childhood Age Continuum (Birth to Age 8)

Kindergarten 5	1st grade 6	2nd grade 7	3rd grade 8 years	Beyond age 8
Teachers use both formal and informal assessments to plan and guide instruction.				
All children should be screened at school entry for vision and hearing needs and checked for immunizations. Some mild disabilities may only become apparent in the school context. Districts and states must by law have sound teacher and parent referral policies, so that children with potential disabilities are referred for in-depth assessment.				
Beginning at age 5, it is possible to use direct measures, including measures of children's early learning, as part of a comprehensive early childhood assessment for monitoring trends. Matrix sampling should be used to ensure technical accuracy and to provide safeguards for individual children. Because of the cost of such an assessment, states or the nation should pick one grade level for monitoring trends in early childhood, most likely kindergarten or first grade.				
teachers, and schools accountable				
			Before age 8, standardized achievement measures are not sufficiently accurate to be used for high-stakes decisions about individual children and schools. Therefore, high-stakes assessments intended for accountability purposes should be delayed until the end of third grade (or preferably fourth grade).	

Photo: Marilyn Nolt

Sample of student work: the North Carolina Grades 1 and 2 Assessment

Purpose 3. Monitoring trends and evaluating programs and services



Definition of purpose. For assessment Purpose 1 and Purpose 2, assessment data were used to make decisions about individual children. For Purpose 3, assessment data are gathered about groups of children in the aggregate and are used by policymakers to make decisions about educational and social programs. In this category, we include two different types of measures, *social indicators*, used to assess the adequacy of services to children or conditions in the environment, and *direct measures of children*, where children themselves are the sources of the data. Examples of social indicators include the percentage of mothers in a state who receive well-baby care, the percentage of 2-year-olds on schedule with immunizations, or the percentage of low-income children who attend quality preschool programs. Direct measures of children's performance could include the degree of language development or familiarity with concepts of print. (For example, does the child come to school knowing how to hold a book and knowing that printed words tell a story?) Such measures, when aggregated for groups of children and used for Purpose 3, could assess the desired outcomes of quality preschool. Note, however, that these assessments are not used to make decisions about the children who participate, but instead are used to evaluate programs.

We have combined within Purpose 3 two closely related uses of aggregate data, *monitoring trends* and *program evaluation*. Large-scale assessment programs such as the National Assessment of Educational Progress (NAEP) serve a monitoring function. Data for the nation and for states are gathered on a regular cycle to

document any changes in levels of student performance. Assessments designed to monitor trends could be used to monitor progress toward Goal 1 or to answer the question, "How is my state doing compared to the United States, another state, or Germany and other industrialized nations?"

Program evaluation refers to large-scale evaluation studies such as the evaluation of preschool, Head Start, or Title I programs. Program evaluations help to document the quality of program delivery and to determine whether programs are effective in achieving intended outcomes. In this sense, the uses of data under Purpose 3 hold programs "accountable" and hold states "accountable" for the adequacy of social conditions and services to young children. However, because the use of data to judge national or state programs entails consequences for the programs rather than for individuals, it is still relatively low-stakes for the individual children, teachers, schools, or local early childhood programs involved. Because of very different implications for technical safeguards, the Goal 1 Early Childhood Assessments Resource Group has drawn a sharp distinction between monitoring and program evaluation uses of data and the high-stakes accountability uses of assessments described in Purpose 4, which entail consequences for individuals.

Audience. Policymakers are the primary audience for aggregate assessment data. Trend data and results of program evaluations are also important to the public and to educators and social service providers with particular responsibility for improving programs. For example, national evaluations of Head Start provide evidence to Congress of the benefits of early educational interventions, which ensures continued funding as well as the establishment of related programs, such as Early Head Start and Even Start. In addition, more detailed evidence gathered as part of Head Start demonstrations and evaluations gives feedback to the system, and can be used for subsequent improvement of the overall Head Start program. For example, early evaluations documented and reinforced the importance of parent involvement in accomplishing and sustaining program goals. Similarly, the data from Goal 1 activities can be used to inform the public regarding the overall status of America's young children, as well as identifying where services are needed to foster children's optimal development.

Technical requirements. Because of their use in making important policy decisions, large-scale assessment data must meet high standards of technical accuracy. For example, if policy changes are going to be made because reading scores have gone up or down, it is essential that the reported change be valid, and not an artifact of measurement error or changes in the test. One of the difficulties, for example, of using teacher opinion surveys to report on kindergartners' readiness for school is that changes over time could be happening because children are becoming more or less ready or because teachers' expectations of readiness vary or are changing. Because of their visibility, state and national assessments also serve important symbolic functions. For example, when the NAEP results are reported, they are often accompanied by sample problems illustrating what students at each age

should know and be able to do. Because teachers and school administrators often make changes in curriculum and instructional strategies in an effort to improve performance on such external assessments, it is important that the NAEP for fourth and eighth graders include challenging open-ended problems, and not just the kinds of questions that lend themselves most easily to multiple-choice formats. Similarly, direct measures of young children should be broadly representative of the five dimensions of early learning and development, and not limited to inappropriate paper-and-pencil tests. In addition, in order to inform public policy adequately, large-scale trend data and evaluation measures should address the conditions of learning—the adequacy of programs, the level of training of caregivers and teachers, the curriculum materials used, and the adequacy of support services—as well as the outcomes of early education and intervention.

Fortunately, the difficulties in measuring young children accurately can be compensated for in Purpose 3 by the aggregate nature of the data. Instead of in-depth assessment of each child needed to ensure reliability and validity for Purpose 2, gathering data from sufficient numbers of children can ensure accuracy for purposes of evaluating programs. *Matrix sampling* is a statistical technique whereby each child participating in the assessment takes only part of the total assessment. Matrix sampling, which is currently used as part of the NAEP design, has two distinct advantages. First, it allows comprehensive coverage of a broad assessment domain without overburdening any one child or student who participates in the assessment. Second, because each student takes only a portion of the total assessment, it is impossible to use the results to make decisions about individual children. This second feature is especially important as a safeguard against misuse of assessment results.

Age continuum. Because of the difficulties in obtaining direct measures of learning with young children, the types of measures that can be included in a monitoring system or evaluation study are very different for children at different ends of the age range from birth to age 8. For children from birth to 2, the only direct measures that are sufficiently accurate to be feasible in a large-scale, every-child data collection effort are measures of physical characteristics such as birthweight. For children in this youngest age range, monitoring systems should focus on the conditions of learning by creating social indicators that track characteristics of families and the adequacy of health and child care services. Important indicators in this earliest age range include percentage of low-birthweight babies or the percentage of 2-year-olds being immunized.

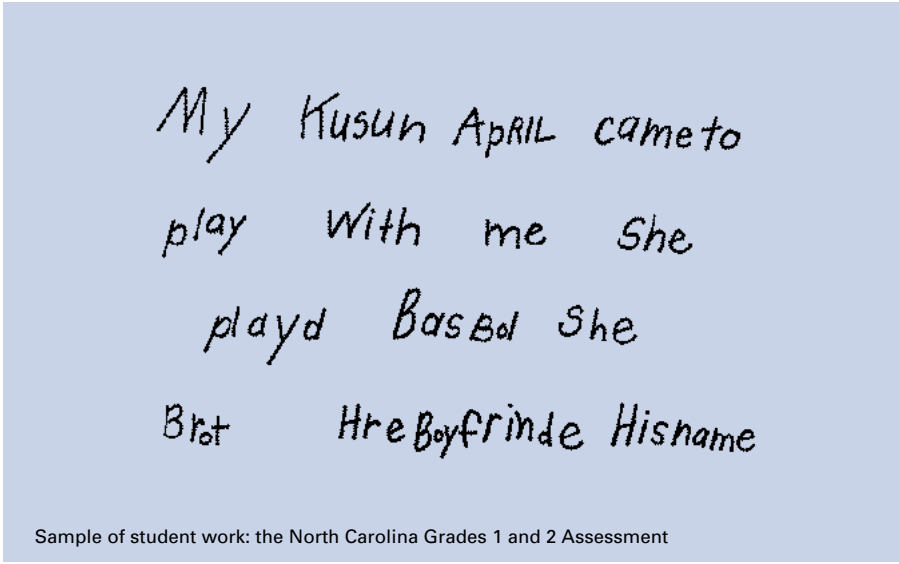
For 3- and 4-year-olds, social indicators that describe the adequacy of services in support of learning and development are presently the preferred mode of assessment. For example, Ohio's annual *Progress Report on Education* reports data on the percentage of 3- and 4-year-olds in poverty who participate in Head Start or preschool. It is also possible to assess learning of 3- and 4-year-olds directly. Although good measures are not readily available off the shelf at present, it is technically possible to construct direct measures of cognitive, language, social, and

motor learning for 3- and 4-year-olds. To avoid overtesting and protect against misuse, these assessments should use matrix sampling procedures. To ensure appropriate and accurate procedures, assessments should be administered to children individually by trained examiners under controlled conditions. Direct measures of learning would be costly to develop and administer, but the information gained would make such efforts feasible if designed as part of targeted national evaluation studies, such as the evaluation of Head Start, Even Start, and Title I in the preschool years. In these studies, data are aggregated to evaluate programs and are not used to make decisions about individual children.

Although direct measures of learning are possible in the context of large-scale program evaluations, it may still be costly and unfeasible to establish a state or national monitoring system to assess 3- or 4-year-olds. The problem would not be just with creating the direct measures themselves, but with the difficulties in locating and sampling all of the 3- or 4-year-olds in a state. Unlike the Head Start example, where the sample could be drawn from those children participating in the program, a state monitoring system would require a household survey and individual assessments for a sample of children in their homes, at a cost that would outweigh potential benefits.

Beginning at age 5, however, it would be possible to administer direct measures of learning outcomes to children in school as part of a monitoring system. For example, the Goals Panel's Goal 1 Resource Group on School Readiness proposed a national Early Childhood Assessment to provide comprehensive information about the status of the nation's children during their kindergarten years. The envisioned assessment would not only address the multiple dimensions of early learning and development, but would also counteract the fallible nature of each data source by collecting information from parents, teachers, and children themselves, through both direct measures and portfolios of classroom work. The five dimensions of early learning suggested by the Resource Group are being used by the National Center for Education Statistics as the framework for developing measures for the National Early Childhood Longitudinal Survey. Although these measures would not be available for widespread use, the insights gained from their development and field testing should be helpful to states trying to develop their own assessments.

Individual states could consider developing an early childhood assessment program for monitoring trends. However, the cost of developing such a system that is both comprehensive and technically sound would be substantial. Therefore, it would be unfeasible to try to collect assessment data at every grade level from kindergarten to Grade 3. Instead, one grade level should be selected for this type of trend data, most likely either kindergarten or Grade 1. A kindergarten-year assessment would have the advantage of being both a culminating measure of the effects of learning opportunities and services available in the years before school and a "baseline" measure against which to compare learning gains by fourth grade. A first grade assessment would be less desirable for monitoring trends because of the



Sample of student work: the North Carolina Grades 1 and 2 Assessment

blurring of preschool and school effects. However, a kindergarten-year assessment would have special sampling problems, because participation in kindergarten is voluntary in many states. At a minimum, accurate interpretation of trend data would require sampling of children in private kindergartens as well as in public schools. In addition, regardless of which grade is used to collect trend data, it would be important to keep track of demographic characteristics, especially first- and second-language status, age, and preschool experience, because changes in these factors have substantial effects and could help in interpreting changes in trend data.

Recommendations for what policymakers can do

1. Before age 5, large-scale assessment systems designed to inform educational and social policy decisions about young children should focus on social indicators that measure the conditions of learning. Direct measures of learning outcomes for 3- and 4-year-olds can be developed and used in large-scale program evaluations, such as Head Start, Even Start, and Title I in the preschool years, but must be administered under controlled conditions and using matrix sampling. Results should not be reported for individual children.

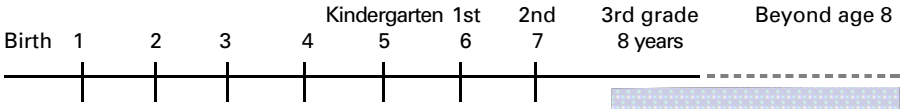
2. Beginning at age 5, it is possible to use direct measures, including measures of children's learning, as part of a comprehensive early childhood system to monitor trends. Matrix sampling procedures should be used to ensure technical accuracy and at the same time protect against the misuse of data to make decisions about individual children. Because such systems are costly to implement, states or the nation should pick one grade level for purposes of monitoring learning trends in early childhood, most likely either kindergarten or first grade.

Photo: Michael Tony

We put books on the table and
nade a Maze for the guinea
Pigs. We put a guinea pig
way in the back of the house.
They went to find the grass
at the Other end of the
house. We were trying to find out
how many Seconds it would take
them to find the grass. It took
Rodney 2 minutes and 20 seconds

Sample of student work: the North Carolina Grades 1 and 2 Assessment

Purpose 4. Assessing academic achievement to hold individual students, teachers, and schools accountable



Before age 8, standardized achievement measures are not sufficiently accurate to be used for high-stakes decisions about individual children and schools. Therefore, high-stakes assessments intended for accountability purposes should be delayed until the end of third grade (or preferably fourth grade).

Definition of purpose. Purpose 4 refers to external examinations, mandated by an authority outside the school, usually the state or school district, and administered to assess academic achievement and to hold students, teachers, and schools accountable for desired learning outcomes. For policymakers, there is a close similarity between the use of assessment data for Purpose 3 and Purpose 4. Both might be used, for example, to report on state and district trends or to compare state and district results to national norms or international benchmarks. However, the important distinction between Purposes 3 and 4 is how individuals who participate in the assessment—teachers and students—are affected by assessment results. Included in this category are external assessments administered nationally or by states and school districts. If results are reported for individual students, classrooms, or schools, then the assessment has much higher stakes than either day-to-day instructional assessments or statewide trend data. Obviously, when assessment results are used to retain students in kindergarten or to award merit pay for teachers, the consequences of assessment are serious. Research evidence shows, however, that merely reporting school results in the newspaper is sufficient to give high stakes to assessment results with accompanying changes produced in instructional practices. Therefore, *the decision to report scores for individual students and schools* places assessments in this "accountability" category, whether or not the assessment is explicitly labeled as an accountability system.

Audience. Policymakers and the general public are, again, the primary audience for accountability data. An expressed intention of school-by-school reporting and reporting of individual student results is to give local constituencies, especially parents, the data they need to be informed about the quality of local schools and to lobby for program improvement.

Sample of student work: the North Carolina Grades 1 and 2 Assessment

Technical requirements. Accountability assessments may be similar in content to assessments used for monitoring trends. Both should be comprehensive measures of important learning goals. At higher grade levels, in fact, some states have school accountability systems that are also used to report state and district trends in achievement. Standards for reliability and validity are more difficult to meet for accountability purposes, however, because *standards for technical accuracy must be met at the lowest unit of reporting*. Thus, individual student scores must be sufficiently reliable, instead of just the state or district mean being reliable. Because

each individual score must be sufficiently reliable and valid, it is not possible to use the aggregation of scores to compensate for inaccuracies in individual measures. Individual-score reporting also precludes the use of matrix sampling to sample an assessment domain broadly. Instead, for fairness reasons, all students must take the same test.

The high-stakes nature of accountability assessments also contributes to their possible inaccuracy. All assessments are fallible and potentially corruptible. Results can be distorted by departures from standardized administration procedures (i.e., allowing more time) or by inappropriate teaching-to-the-test (i.e., giving practice on questions that closely resemble the assessment). These practices are documented to occur more frequently when the results of testing have high-stakes consequences for students or teachers. Although some educators may be motivated by personal gain to coach their students or to change answers, widespread practices that undermine the integrity of results are more likely to occur because a test is seen as professionally indefensible, because it is unfair to children, takes time away from teaching, or diverts attention from important learning goals.

Age continuum. Direct measures of learning outcomes are fraught with error throughout the entire early childhood age span. Such errors have very different consequences in an accountability context than in classroom contexts, where teachers are constantly learning new things about each child. Although standardized measures of children's physical, social, emotional, and cognitive development could be constructed and administered for purposes of program evaluation and monitoring trends—because data aggregation would provide both safeguards and improved accuracy—such assessments cannot be made sufficiently reliable and fair to be used for high-stakes decisions about individual children and schools.

Recommendations for what policymakers can do

1. Before age 8, standardized achievement measures are not sufficiently accurate to be used for high-stakes decisions about individual children and schools. Therefore, high-stakes assessments intended for accountability purposes should be delayed until the end of third grade (or preferably fourth grade).

2. Although it is not technically defensible for states or districts to administer formal, standardized measures to hold first and second graders to grade-level standards, policymakers have a legitimate concern that 3rd grade is "too late" to identify children who are falling behind. As suggested under Purpose 1, policymakers at the state and district level could reasonably require that teachers and the schools have procedures in place to monitor student progress using instructionally relevant assessments, and that schools have a plan for providing intensified special help if children are having difficulty, especially in learning to read.

Combining Assessment Purposes

There is a natural tendency for policymakers and educators to want to use assessment data for more than one purpose. The cost of developing new assessments would be better justified if the results could be used for multiple purposes, and if teachers and children go to the trouble of participating in an assessment, it would be desirable to get as much use from the data as possible. Many parents, teachers, and policymakers also want a combined system so that individual student results can be compared to standards set by the state or district. However, these desires for efficient use of assessment results must be weighed against the abuses that have occurred in the past when instruments designed for one purpose were misused for another.

Often, it is a mistake to combine purposes. This is true either because the different purposes require different assessment *content* or because the *technical requirements* for each purpose are quite different. In the examples that follow, we consider the combinations of purposes that have most often occurred in practice, either in early childhood settings or in state assessment programs. In the first case, educators and policymakers frequently confuse the use of instruments intended for Purpose 1 and Purpose 2, thinking that it is legitimate to do so because both involve assessments of individual children. They are not aware, however, that the two purposes require different content as well as different levels of technical rigor.

Similarly, it seems reasonable to use the same assessments to serve Purposes 1, 3, and 4 on the grounds that all three involve measures of learning outcomes. However, reporting individual student and school-level data for accountability purposes (Purpose 4) requires a higher level of technical accuracy than the other two purposes, a level of accuracy that cannot be attained in large-scale programs for children younger than age 8. Therefore, the Resource Group has made quite different recommendations before and after Grade 3 regarding the feasibility of including accountability uses of assessment data.

Individual assessments, Purposes 1 and 2. In the past, screening measures intended as a first step in referral for special-needs identification have been misused for instructional purposes. For example, screening instruments designed to resemble short-form IQ tests have been used inappropriately to plan instruction or to hold children out of kindergarten. Although it would be possible, in theory, to develop assessments that could be used legitimately for both classroom assessment and screening for special needs (Purposes 1 and 2), extensive investment would be required to develop both curricularly relevant assessment content and empirical norms for evaluating disability.

To support teaching and learning (Purpose 1), assessment tasks should be as closely tied to the local preschool or primary curriculum as possible. For Purpose 2, when clinicians are trying to make inferences about *ability* to learn and/or the existence of a possible *disability,* IQ tests and other developmental measures have traditionally been designed to be as "curriculum free" as possible. The intention is

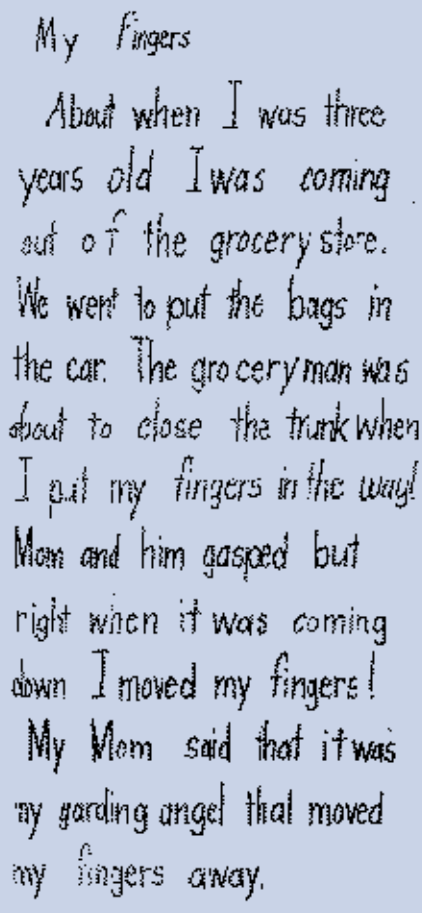
to use the most generic tasks possible, so that all children from a wide variety of backgrounds will be equally familiar with the content of the assessment. Of course, this has not always worked even when seemingly familiar content was used; hence the problems of cultural bias.

An alternative method of assessment for special-needs identification would be to use dynamic assessment, where ability to learn is evaluated over time by providing focused learning opportunities interactively with assessment. Dynamic assessment techniques have not yet been sufficiently developed to permit their dissemination for widespread use. Even school psychologists and other specialists would need extensive training to use dynamic assessment with curriculum-aligned assessment tasks. We should also note that assessment materials intended for use in making special education placement decisions would require normative data and an empirical basis to support interpreting low performance as evidence of a disability, and would have to meet the more stringent reliability and validity standards for Purpose 2. In the meantime, the most appropriate policies are those that prevent the misuse of existing instruments.

Assessments of learning outcomes for Grade 3 and above, Purposes 1, 3, and 4. At higher grade levels, states have attempted to develop measures of academic outcomes that could be used for individual instructional decisions, reporting of state-level achievement trends, and school accountability. Kentucky's use of classroom portfolios for school and state reporting is one such example. Use of assessments for multiple purposes requires significant investment of resources to ensure that the technical requirements for *each* purpose are satisfied. There may also be some sacrifices required from the design that would be optimal for each purpose separately. In the Kentucky example, the intention to use results for school and state-level comparisons requires that the tasks or entries in the portfolios be the same for a given grade and subject matter. Such standardization of curricular expectations would not be possible nationally or in states without a state curriculum. Use for accountability purposes also requires standardization of scoring across schools and rigorous external checks to make sure that the data being aggregated from classrooms are comparable. There are many benefits to this articulated, multipurpose assessment system, but it also requires substantial investment of resources.

Assessments of learning outcomes before Grade 3, Purposes 1 and 3. Because of the inherent difficulties of assessing young children accurately, and the heightened problems and technical requirements of high-stakes testing, the Resource Group has recommended against accountability uses of assessment data before the end of Grade 3. For the same reasons, it is unworkable to attempt to combine assessments for Purpose 1 and Purpose 4 for early grade levels. Assessments could not at the same time be flexible and informal enough to be useful to teachers in day-to-day teaching and learning and still meet the technical requirements of reliability, standardization, comparability, validity, and fairness that must be satisfied for accountability reporting.

States considering early childhood assessments to monitor trends (Purpose 3, a low-stakes type of program accountability) could, however, work to ensure that the content of assessments used for Purpose 1 is closely aligned with the content of the statewide assessment. For example, as part of developing continua of proficiencies in the early grades that lead to attainment of state performance standards in Grade 3 or Grade 4, states could develop model instructional units with accompanying assessments to be used as part of the learning process. Such materials could be made available to local districts to aid in curriculum improvement and staff development, but would not be formally administered as part of a state assessment. Because of differences in technical requirements, the exact same assessment would not be used for Purpose 1 and Purpose 3, but the two types of assessments could be developed in parallel so that they would be conceptually compatible and mutually supportive.



My fingers

About when I was three years old I was coming out of the grocery store. We went to put the bags in the car. The grocery man was about to close the trunk when I put my fingers in the way! Mom and him gasped but right when it was coming down I moved my fingers! My Mom said that it was my garding angel that moved my fingers away.

Sample of student work: the North Carolina Grades 1 and 2 Assessment

Conclusions

Assessment of young children is important both to support the learning of each individual child and to provide data—at the district, state, and national level—for improving services and educational programs. At the level of the individual child, teaching and assessment are closely linked. Finding out, on an ongoing basis, what a child knows and can do, helps parents and teachers decide how to pose new challenges and provide help with what the child has not yet mastered. Teachers also use a combination of observation and formal assessments to evaluate their own teaching and make improvements. At the policy level, data are needed about the preconditions of learning—such as the adequacy of health care, child care, and preschool services. Direct measures of children's early learning are also needed to make sure that educational programs are on track in helping students reach high standards by the end of third grade.

Assessing young children accurately is much more difficult than for older students and adults, because of the nature of early learning and because the language skills needed to participate in formal assessments are still developing. Inappropriate testing of young children has sometimes led to unfair and harmful decisions. Such testing abuses occur primarily for one of two reasons: either a test designed for one purpose is improperly used for another purpose, or testing procedures appropriate for older children are used inappropriately with younger children. In making its recommendations, the Resource Group has emphasized how technical requirements for assessments must be tailored to each assessment purpose, and we have tried to explain how the increasing reliability and validity of measurement for ages from birth to age 8 should guide decisions about what kinds of assessments can be administered accurately at each age.

Four categories of assessment purpose were identified, with accompanying recommendations for educators and policymakers:

- **Assessing to promote children's learning and development.** The most important reason for assessing young children is to help them learn. To this end, assessments should be closely tied to preschool and early-grades curriculum, and should be a natural part of instructional activities. Policymakers should support the development or provision of assessment materials, to be used instructionally, that exemplify important and age-appropriate learning goals. States should also support professional development to help teachers learn to use benchmark information to extend children's thinking.

- **Assessing to identify children for health and special services.** Screening or a referral procedure should be in place to ensure that children suspected of having a health or learning problem are referred for in-depth evaluation. Given the potential for misuse of cognitive screening measures, states that mandate screening tests should monitor how they are used and should take extra steps to avoid inappropriate uses. IQ-like tests should not be used to

exclude children from school or to plan instruction. Often, the need for costly assessments could be eliminated if intensive language and literacy programs were more broadly available for all of the groups deemed educationally at-risk, e.g., children living in poverty, children with mild cognitive and language disabilities, and children with early reading difficulties.

- **Assessing to monitor trends and evaluate programs and services.** The kinds of assessment that teachers use in preschool and the early grades to monitor children's learning are not sufficiently reliable or directly comparable for uses outside the classroom. Before age 5, assessment systems designed to gather data at the state or national level should focus on social indicators that describe the conditions of learning, e.g., the percentage of low-income children who attend quality preschool programs. Beginning at age 5, it is possible to develop large-scale assessment systems to report on trends in early learning, but matrix sampling should be used to ensure technical accuracy and at the same time protect individual children from test misuse.

- **Assessing academic achievement to hold individual students, teachers, and schools accountable.** *There should be no high-stakes accountability testing of individual children before the end of third grade.* This very strong recommendation does not imply that members of the Resource Group are against accountability or against high standards. In fact, instructionally relevant assessments designed to support student learning should reflect a clear continuum of progress in Grades K, 1, and 2 that leads to expected standards of performance for the third and fourth grades. Teachers should be accountable for keeping track of how well their students are learning and for responding appropriately, but the technology of testing is not sufficiently accurate to impose these decisions using an outside assessment.

Congress charged the Goals Panel advisors to offer "clear guidelines regarding the nature, functions, and uses of early childhood assessments." In examining current trends in state and local policies, we found numerous efforts to guard against testing misuses of the past, as well as positive efforts to devise standards and assessments that would clearly document children's learning. We hope that these recommendations and principles will be useful to educators and parents, as well as to state policymakers who hold the authority for determining testing policies. Ultimately, our goal is to set high expectations for early learning and development, to make sure that no child who falls behind goes unnoticed, and at the same time to help parents and the public understand how varied are the successful paths of early learning, depending on the rate of development, linguistic and cultural experiences, and community contexts.

Glossary

Accountability: The concept of trying to hold appropriate parties accountable for their performance; in education these are usually administrators, teachers, and/or students. Beyond fiscal accountability, this concept currently means responsibility for student academic performance, usually by publicly reporting student achievement data (often test scores). Accountability mechanisms vary among states and local districts in the types of school and student data that are used and in the degree to which rewards, sanctions, or other consequences are attached to performance.

Assessment: The process of collecting data to measure the performance or capabilities of a student or group. Paper-and-pencil tests of students' knowledge are a common form of assessment, but data on student attendance or homework completion, records of informal adult observations of student proficiency, or evaluations of projects, oral presentations, or other forms of problem-solving may also be assessments.

Child Find programs: Organized efforts by health, welfare, and education agencies to locate and identify children in need of special education services.

Development: Growth or maturation that occurs primarily because of the emergence of underlying biological patterns or preconditions. The terms *development* and *learning* are distinguished by the presumption that one is caused by genetics and the other by experience. However, it is known that development can be profoundly affected by environmental conditions.

Developmental assessment: Measurement of a child's cognitive, language, knowledge, and psychomotor skills in order to evaluate development in comparison to children of the same chronological age.

Developmental continuum: A continuum that describes typical milestones in children's growth and emerging capabilities according to age.

Dynamic assessment: An interactive mode of assessment used to evaluate a child's ability to learn by providing a structured learning situation, observing how the child performs, and evaluating how well the child is able to learn new material under various conditions of supported learning.

Early childhood: The stage of life from birth through age 8.

Formal assessment: A systematic and structured means of collecting information on student performance that both teachers and students recognize as an assessment event.

High-stakes assessment: Assessments that carry serious consequences for students or for educators. Their outcomes determine such important things as promotion to the next grade, graduation, merit pay for teachers, or school rankings reported in the newspaper.

Informal assessment: A means of collecting information about student performance in naturally occurring circumstances, which may not produce highly accurate and systematic results, but can provide useful insights about a child's learning.

Large-scale assessment: Standardized tests and other forms of assessment designed to be administered to large groups of individuals under prescribed conditions to provide information about performance on a standardized scale so that results for districts, states, or nations can be fairly compared.

Learning: Acquiring of knowledge, skill, ways of thinking, attitudes, and values as a result of experience.

Matrix sampling: A way to select a subset of all the students to be tested and subsets of various parts of a test so that each student takes only a portion of the total assessment, but valid conclusions can be drawn about how all students would have performed on the entire test.

Norms: Statistics or data that summarize the test performance of specified groups such as test-takers of various ages or grades.

Normal variation: Refers to the range of performance levels that, in addition to the average (or mean) performance, is typical for children of a specific age or grade.

Observation: A systematic way to collect data by watching or listening to children during an activity.

Portfolio: An organized and purposeful collection of student work and self-assessments collected over time to demonstrate student learning. A *portfolio assessment* is the process of evaluating student achievement based on portfolios.

Readiness test: A test used to evaluate a student's preparedness for a specific academic program.

Reliability: The degree to which a test or assessment measures consistently across different instances of measurement—for example, whether results are consistent across raters, times of measurement, or sets of test items.

Screening: Selecting individuals on a preliminary test who are in need of more thorough evaluation.

Screening test: A test used as a first step in identifying children who may be in need of special services. If a potential problem is suggested by the results of a screening test, then a child should be referred for a more complete assessment and diagnosis.

Social indicator: A statistic (usually not a student test result) used to report on a societal condition, such as the rate of infant mortality, teen pregnancy, or school dropouts.

Special education: As defined by regulations of the Individuals with Disabilities Education Act, special education is the specially designed instruction that public schools are required to offer either in a separate or regular classroom to meet the unique needs of a child with a disability.

Standardized test or assessment: Standardization refers to a set of consistent procedures for administering and scoring a test or assessment. Standardization is necessary to make test scores comparable across individuals.

Test: A formal procedure for eliciting responses so as to measure the performance and capabilities of a student or group.

Validity: The accuracy of a test or assessment in measuring what it was intended to measure. Validity is determined by the extent to which interpretations and decisions based on test scores are warranted and supported by independent evidence.

Sources

McDonnell, L.M., McLaughlin, M.J., & Morison, P. (Eds.). (1997). *Educating one and all: Students with disabilities and standards-based reform*. Washington, DC: National Academy Press.

McLaughlin, M.W., & Shepard, L.A. (1995). *Improving education through standards-based reform*. Stanford, CA: National Academy of Education.

National Association for the Education of Young Children. (1988). NAEYC position statement on standardized testing of young children 3 through 8 years of age. *Young Children* 43(3): 42–47.

Bibliography

Bredekamp, S., & Copple, C. (Eds.). (1997). *Developmentally appropriate practice in early childhood programs* (Rev. ed.). Washington, DC: National Association for the Education of Young Children.

British Columbia Ministry of Education. (1991). Principles of assessment and evaluation. Primary Program Foundation Document. Available in adapted form in Nebraska Department of Education. (1994). *Nebraska/Iowa Primary Program: Growing and learning in the heartland* (pp. 95–108). Lincoln, NE: Author.

California Department of Education, Child Development Division. (1992, July). *Appropriate assessment practices for young children*. [Program Advisory]. Sacramento: Author.

Federal Interagency Forum on Child and Family Statistics. (1997). *America's children: Key national indicators of well-being*. Washington, DC: U.S. Government Printing Office.

Gredler, G.R. (1992). *School readiness: Assessment and educational issues*. Brandon, VT: Clinical Psychology Publishing Co.

Greenspan, S.I., & Meisels, S.J. (1996). Toward a new vision for the developmental assessment of infants and young children. In S.J. Meisels & E. Fenichel (Eds.), *New visions for the developmental assessment of infants and young children*. Washington, DC: ZERO TO THREE: The National Center for Infants, Toddlers, and Families.

High/Scope Educational Research Foundation. (1992). *High/Scope Child Observation Record (COR) for ages 2½–6*. Ypsilanti, MI: High/Scope Press.

Hills, T.W. (1997). Finding what is of value in programs for young children and their families. In C. Seefeldt & A. Galper (Eds.), *Continuing issues in early childhood education* (2nd ed., pp. 293–313). Upper Saddle River, NJ: Prentice-Hall, Inc.

Hills, T.W. (1992). Reaching potentials through appropriate assessments. In S. Bredekamp & T. Rosegrant (Eds.), *Reaching potentials: Appropriate curriculum and assessment for young children* (pp. 43–63). Washington, DC: National Association for the Education of Young Children.

Kagan, S.L., Moore, E., & Bredekamp, S. (Eds.). (1995, June). *Reconsidering children's early development and learning: Toward common views and vocabulary*. Goal 1 Technical Planning Group Report 95–03. Washington, DC: National Education Goals Panel.

Kagan, S.L., Rosenkoetter, S., & Cohen, N. (1997). *Considering child-based results for young children: Definitions, desirability, feasibility, and next steps*. Based on Issues Forums on Child-Based Results, sponsored by the W.W. Kellogg Foundation, the Carnegie Corporation of New York, and Quality 2000: Advancing Early Care and Education. New Haven, CT: Yale Bush Center in Child Development and Social Policy.

Langhorst, B.H. (1989, April). *A consumer's guide: Assessment in early childhood education*. Portland, OR: Northwest Regional Educational Laboratory.

Meisels, S.J. (1994). Designing meaningful measurements for early childhood. In B.L. Mallory & R.S. New (Eds.), *Diversity in early childhood education: A call for more inclusive theory, practice, and policy* (pp. 205–225). New York: Teachers College Press.

Meisels, S.J. (1989). High-stakes testing in kindergarten. *Educational Leadership* 46(7): 16–22.

Meisels, S.J. (1987). Uses and abuses of developmental screening and school readiness testing. *Young Children 42*: 4–6, 68–73.

Meisels, S.J., with Atkins-Burnett, S. (1994). *Developmental screening in early childhood: A*

guide (4th ed.). Washington, DC: National Association for the Education of Young Children.

Meisels, S.J., & Fenichel, E. (Eds.). (1996). *New visions for the developmental assessment of infants and young children*. Washington, DC: ZERO TO THREE: National Center for Infants, Toddlers, and Families.

Meisels, S.J., Jablon, J.R., Marsden, D.B., Dichtelmiller, M.L., & Dorfman, A.B. (1994). *The Work Sampling System*. Ann Arbor, MI: Rebus, Inc.

Meisels, S.J., Marsden, D.B., Wiske, M.S., & Henderson, L.W. (1997). *The Early Screening Inventory* (Rev. ed.). [ESI=B7R]. Ann Arbor, MI: Rebus, Inc.

Meisels, S.J., & Provence, S. (1989). *Screening and assessment: Guidelines for identifying young disabled and developmentally vulnerable children and their families*. Washington, DC: National Center for Clinical Infant Programs.

Michigan State Board of Education, Early Childhood Education & Parenting Office. (1992, April). *Appropriate assessment of young children*. Lansing: Michigan Department of Education.

Minnesota Department of Education. (1990). *Model learner outcomes for early childhood education*. St. Paul: Author.

National Association for the Education of Young Children and National Association of Early Childhood Specialists in State Departments of Education. (1991). Guidelines for appropriate curriculum content and assessment in programs serving children ages 3 through 8. *Young Children 46*(1): 21–38.

National Association for the Education of Young Children. (1988). NAEYC position statement on standardized testing of young children 3–8 years of age. *Young Children 43*(3): 42–47.

National Education Goals Panel. (1997, January). *Getting a good start in school*. Washington, DC: U.S. Government Printing Office.

National Education Goals Panel. (1997, October). *Special early childhood report 1997*. Washington, DC: U.S. Government Printing Office.

National Forum on Education Statistics. (1994). *A statistical agenda for early childhood care and education: Addendum to "A Guide to Improving the National Education Data System."* Adopted by the National Forum on Education Statistics, January 1994.

Neisworth, J.T. (1993). Assessment: DEC recommended practices. In *DEC recommended practices: Indicators of quality in programs for infants and young children with special needs and their families*. (see EC 301 933).

Perrone, V. (1991). On standardized testing: A position paper of the Association for Childhood Education International. *Childhood Education 67*(3): 131–142.

Puckett, M.B., & Black, J.K. (1994). *Authentic assessment of the young child: Celebrating development and learning*. New York: Merrill, an imprint of Macmillan College Publishing Company.

Shepard, L.A. (1994). The challenges of assessing young children appropriately. *Phi Delta Kappan* 76(3): 206–213.

Shepard, L.A. (1997). Children not ready to learn? The invalidity of school readiness testing. *Psychology in the Schools 34*(2): 85–97.

Shepard, L.A. (1991). The influence of standardized tests on the early childhood curriculum, teachers, and children. In B. Spodek & O.N. Saracho (Eds.), *Yearbook in early childhood education* (Vol. 2). New York: Teachers College Press.

Goal 1 Advisors to the National Education Goals Panel

Technical Planning Group on Readiness for School

Leader: Sharon Lynn Kagan, Yale University

Sue Bredekamp, National Association for the Education of Young Children
M. Elizabeth Graue, University of Wisconsin
Luís Laosa, Educational Testing Service
Samuel Meisels, University of Michigan
Evelyn Moore, National Black Child Development Institute
Lucile Newman, Brown University
Lorrie Shepard, University of Colorado
Valora Washington, The Kellogg Foundation
Nicholas Zill, Westat, Inc.

Goal 1 Early Childhood Assessments Resource Group

Leaders: Sharon Lynn Kagan, Yale University
Lorrie Shepard, University of Colorado
Sue Bredekamp, National Association for the Education of Young Children
Edward Chittenden, Educational Testing Service
Harriet Egertson, Nebraska State Department of Education
Eugene García, University of California, Berkeley
M. Elizabeth Graue, University of Wisconsin
Kenji Hakuta, Stanford University
Carollee Howes, University of California, Los Angeles
Annemarie Palincsar, University of Michigan
Tej Pandey, California State Department of Education
Catherine Snow, Harvard University
Maurice Sykes, District of Columbia Public Schools
Valora Washington, The Kellogg Foundation
Nicholas Zill, Westat, Inc.

Goal 1 Ready Schools Resource Group

Leaders: Asa Hilliard, Georgia State University
Sharon Lynn Kagan, Yale University
Barbara Bowman, Erikson Institute
Cynthia Brown, Council of Chief State School Officers
Fred Brown, Boyertown Elementary School, Boyertown, Pennsylvania
Linda Espinosa, University of Missouri
Donna Foglia, Norwood Creek School, San Jose, California
Peter Gerber, MacArthur Foundation
Sarah Greene, National Head Start Association
Judith Heumann, U.S. Department of Education
Mogens Jensen, National Center for Mediated Learning
Lilian Katz, ERIC Clearinghouse for Elementary and Early Childhood Education
Michael Levine, Carnegie Corporation of New York
Evelyn Moore, National Black Child Development Institute
Tom Schultz, National Association of State Boards of Education
Barbara Sizemore, DePaul University
Robert Slavin, Johns Hopkins University

Typography and design by the U.S. Government Printing Office.
Editorial assistance provided by Scott Miller, Editorial Experts, Inc.

THE NATIONAL EDUCATION GOALS

READY TO LEARN

MATHEMATICS AND SCIENCE

SCHOOL COMPLETION

ADULT LITERACY AND LIFELONG LEARNING

STUDENT ACHIEVEMENT AND CITIZENSHIP

SAFE, DISCIPLINED, AND ALCOHOL- AND DRUG-FREE SCHOOLS

TEACHER EDUCATION AND PROFESSIONAL DEVELOPMENT

PARENTAL PARTICIPATION

NATIONAL EDUCATION GOALS PANEL
1255 22nd Street, N.W., Suite 502
Washington, DC 20037
202–724–0015 • FAX 202–632–0957
http://www.negp.gov
E-mail: NEGP@goalline.org