The Challenges of Assessing Young Children Appropriately

In the past decade, testing of 4-, 5-, and 6-year-olds has been excessive and inappropriate. Given this history of misuse, Ms. Shepard maintains, the burden of proof must rest with assessment advocates to demonstrate the usefulness of assessment and to ensure that abuses will not recur.

By Lorrie A. Shepard

PROPOSALS to "assess" young children are likely to be met with outrage or enthusiasm, depending on one's prior experience and one's image of the testing involved. Will an inappropriate paper-and-pencil test be used to keep some 5-year-olds out of school? Or will the assessment, implemented as an ordinary part of good instruction, help children learn? A governor advocating a test for every preschooler in the nation may have in mind the charts depicting normal growth in the pediatrician's office. Why shouldn't parents have access to similar measures to monitor their child's cognitive and academic progress? Middle-class parents, sanguine about the use of test scores to make college-selection decisions, may be eager to have similar tests determine their child's entrance into preschool or kindergarten. Early childhood experts, however, are more likely to respond with alarm because they are more familiar with the complexities of defining and measuring development and learning in young children and because they are more aware of the widespread abuses of readiness testing that occurred in the 1980s.

Given a history of misuse, it is impossible to make positive recommendations about how assessments could be used to monitor the progress of individual children or to evaluate the quality of educational programs without offering assurances that the abuses will not recur. In what follows, I summarize the negative history of standardized testing of young children in order to highlight the transformation needed in both the substance and purposes of early childhood assessment. Then I explain from a measurement perspective how the features of an assessment must be tailored to match the purpose of the assessment. Finally, I describe differences in what assessments might look like when they are used for purposes of screening for handicapping conditions, supporting instruction, or monitoring state and national trends.

Note that I use the term *test* when referring to traditional, standardized developmental and pre-academic measures and the term *assessment* when referring to more developmentally appropriate procedures for observing and evaluating young children. This is a semantic trick that plays on the different connotations of the two terms. Technically, they mean the same thing. Tests, as defined by the *Standards for Educational and Psychological Testing*, have always included systematic observations of behavior, but our experience is with tests as more formal, one-right-answer instruments used to rank and sort individuals. As we shall see, assessments might be standardized, involve paper-and-pencil responses, and so on, but in contrast to traditional testing, "assessment" implies a substantive focus on student learning for the purpose of effective intervention. While *test* and *assessment* cannot be reliably distinguished technically, the difference between these two terms as they have grown up in common parlance is of symbolic importance. Using the term *assessment* presents an opportunity to step away from past practices and ask why we should try to measure what young children know and can do. If there are legitimate purposes for gathering such data, then we can seek the appropriate content and form of assessment to align with those purposes.

Negative History of Testing Young Children

In order to understand the negative history of the standardized testing of young children in the past decade, we need to understand some larger shifts in curriculum and teaching practices. The distortion of the curriculum of the early grades dur-

LORRIE A. SHEPARD is a professor of education at the University of Colorado, Boulder. She is past president of the National Council on Measurement in Education, past vice president of the American Educational Research Association, and a member of the National Academy of Education. She wishes to thank Sharon Lynn Kagan, M. Elizabeth Graue, and Scott F. Marion for their thoughtful suggestions on drafts of this article.

Phi Delta Kappan, 76: 206-212. 1994.

ing the 1980s is now a familiar and well-documented story. Indeed, negative effects persist in many school districts today.

Although rarely the result of conscious policy decisions, a variety of indirect pressures — such as older kindergartners, extensive preschooling for children from affluent families, parental demands for the teaching of reading in kindergarten, and accountability testing in higher grades — produced a skill-driven kindergarten curriculum. Because what once were first-grade expectations were shoved down to kindergarten, these shifts in practice were referred to as the "escalation of curriculum" or "academic trickle-down." The result of these changes was an aversive learning environment inconsistent with the learning needs of young children. Developmentally inappropriate instructional practices, characterized by long periods of seatwork, high levels of stress, and a plethora of fill-in-the-blank worksheets, placed many children at risk by setting standards for attention span, social maturity, and academic productivity that could not be met by many normal 5-year-olds.

Teachers and school administrators responded to the problem of a kindergarten environment that was increasingly hostile to young children with several ill-considered policies: raising the entrance age for school, instituting readiness screening to hold some children out of school for a year, increasing retentions in kindergarten, and creating two-year programs with an extra grade either before or after kindergarten. These policies and practices had a benign intent: to protect children from stress and school failure. However, they were ill-considered because they were implemented without contemplating the possibility of negative side effects and without awareness that retaining some children and excluding others only exacerbated the problems by creating an older and older population of kindergartners.[1] The more reasonable corrective for a skill-driven curriculum at earlier and earlier ages would have been curriculum reform of the kind exemplified by the recommendations for developmentally appropriate practices issued by the National Association for the Education of Young Children (NAEYC), the nation's largest professional association of early childhood educators.[2]

The first response of many schools, however, was not to fix the problem of inappropriate curriculum but to exclude those children who could not keep up or who might be harmed. Readiness testing was the chief means of implementing policies aimed at removing young children from inappropriate instructional programs. Thus the use of readiness testing increased dramatically during the 1980s and continues today in many school districts.[3]

Two different kinds of tests are used: developmental screening measures, originally intended as the first step in the evaluation of children for potential handicaps; and pre-academic skills tests, intended for use in planning classroom instruction.[4] The technical and conceptual problems with these tests are numerous.[5] Tests are being used for purposes for which they were never designed or validated. Waiting a year or being placed in a two-year program represents a dramatic disruption in a child's life, yet not one of the existing readiness measures has sufficient reliability or predictive validity to warrant making such decisions.

Developmental and pre-academic skills tests are based on outmoded theories of aptitude and learning that originated in the 1930s. The excessive use of these tests and the negative consequences of being judged unready focused a spotlight on the tests' substantive inadequa-

Illustration by Kay Salem

cies. The widely used Gesell Test is made up of items from old I.Q. tests and is indistinguishable statistically from a measure of I.Q.; the same is true for developmental measures that are really short-form I.Q. tests. Assigning children to different instructional opportunities on the basis of such tests carries forward nativist assumptions popular in the 1930s and 1940s. At that time, it was believed that I.Q. tests could accurately measure innate ability, unconfounded by prior learning experiences. Because these measured "capacities" were thought to be fixed and unalterable, those who scored poorly were given low-level training consistent with their supposedly limited potential. Tests of academic content might have the promise of being more instructionally relevant than disguised I.Q. tests, but, as Anne Stallman and David Pearson have shown, the decomposed and decontextualized prereading skills measured by traditional readiness tests are not compatible with current research on early literacy.[6]

Readiness testing also raises serious equity concerns. Because all the readiness measures in use are influenced by past opportunity to learn, a disproportionate number of poor and minority children are identified as unready and are excluded from school when they most need it. Thus children without preschool experience and without extensive literacy experiences at home are sent back to the very environments that caused them to score poorly on readiness measures in the first place. Or, if poor and minority children who do not pass the readiness tests are admitted to the school but made to spend an extra year in kindergarten, they suffer disproportionately the stigma and negative effects of retention.

The last straw in this negative account of testing young children is the evidence that fallible tests are often followed by ineffective programs. A review of controlled studies has shown no academic benefits from retention in kindergarten or from extra-year programs, whether developmental kindergartens or transitional first grades. When extra-year children finally get to first grade, they do not do better on average than equally "unready" children who go directly on to first grade.[7] However, a majority of children placed in these extra-year programs do experience some short- or long-term trauma, as reported by their parents.[8] Contrary to popular belief that kindergarten children are "too young to notice" retention, most of them know that they are not making "normal" progress, and many continue to make reference to the decision years later. "If I hadn't spent an extra year in kindergarten, I would be in __ grade now." In the face of such evidence, there is little wonder that many early childhood educators ask why we test young children at all.

Principles for Assessment And Testing

The NAEYC and the National Association of Early Childhood Specialists in State Departments of Education have played key roles in informing educators about the harm of developmentally inappropriate instructional practices and the misuse of tests. In 1991 NAEYC published "Guidelines for Appropriate Curriculum Content and Assessment in Programs Serving Children Ages 3 Through 8."[9] Although the detailed recommendations are too numerous to be repeated here, a guiding principle is that *assessments should bring about benefits for children, or data should not be collected at all.* Specifically, assessments "should not be used to recommend that children stay out of a program, be retained in grade, or be assigned to a segregated group based on ability or developmental maturity."[10] Instead, NAEYC acknowledges three legitimate purposes for assessment: 1) to plan instruction and communicate with parents, 2) to identify children with special needs, and 3) to evaluate programs.

Although NAEYC used *assessment* in its "Guidelines," as I do, to avoid associations with inappropriate uses of tests, both the general principle and the specific guidelines are equally applicable to formal testing. In other words, tests should not be used if they do not bring about benefits for children. In what follows I summarize some additional principles that can ensure that assessments (and tests) are beneficial and not harmful. Then, in later sections, I consider each of NAEYC's recommended uses for assessment, including national, state, and local needs for program evaluation and accountability data.

I propose a second guiding principle for assessment that is consistent with the NAEYC perspective. *The content of assessments should reflect and model progress toward important learning goals.* Conceptions of what is important to learn should take into account both physical and social/emotional development as well as cognitive learning. For most assessment purposes in the cognitive domain, content should be congruent with subject matter in emergent literacy and numeracy. In the past, developmental measures were made as "curriculum free" or "culture free" as possible in an effort to tap biology and avoid the confounding effects of past opportunity to learn. Of course, this was an impossible task because a child's ability to "draw a triangle" or "point to the ball on top of the table" depends on prior experiences as well as on biological readiness. However, if the purpose of assessment is no longer to sort students into programs on the basis of a one-time measure of ability, then it is possible to have assessment content mirror what we want children to learn.

A third guiding principle can be inferred from several of the NAEYC guidelines. *The methods of assessment must be appropriate to the development and experiences of young children.* This means that — along with written products — observation, oral readings, and interviews should be used for purposes of assessment. Even for large-scale purposes, assessment should not be an artificial and decontextualized event; instead, the demands of data collection should be consistent with children's prior experiences in classrooms and at home. Assessment practices should recognize the diversity of learners and must be in accord with children's language development — both in English and in the native languages of those whose home language is not English.

A fourth guiding principle can be drawn from the psychometric literature on test validity. *Assessments should be tailored to a specific purpose.* Although not stated explicitly in the NAEYC document, this principle is implied by the recommendation of three sets of guidelines for three separate assessment purposes.

Matching the Why and How Of Assessment

The reason for any assessment — i.e., how the assessment information will be used — affects the substance and form of

The intended use of an assessment will determine the need for normative information or other means to support the interpretation of results.

the assessment in several ways. First, the degree of technical accuracy required depends on use. For example, the identification of children for special education has critical implications for individuals. Failure to be identified could mean the denial of needed services, but being identified as in need of special services may also mean removal from normal classrooms (at least part of the time) and a potentially stigmatizing label. A great deal is at stake in such assessment, so the multifaceted evaluation employed must have a high degree of reliability and validity. Ordinary classroom assessments also affect individual children, but the consequences of these decisions are not nearly so great. An inaccurate assessment on a given day may lead a teacher to make a poor grouping or instructional decision, but such an error can be corrected as more information becomes available about what an individual child "really knows."

Group assessment refers to uses, such as program evaluation or school accountability, in which the focus is on group performance rather than on individual scores. Although group assessments may need to meet very high standards for technical accuracy, because of the high stakes associated with the results, the individual scores that contribute to the group information do not have to be so reliable and do not have to be directly comparable, so long as individual results are not reported. When only group results are desired, it is possible to use the technical advantages of matrix sampling — a technique in which each participant takes only a small portion of the assessment — to provide a rich, in-depth assessment of the intended content domain without overburdening any of the children sampled. When the "group" is very large, such as all the fourth-graders in a state or in the nation, then assessing a representative sample will produce essentially the same results for the group average as if every student had been assessed.

Purpose must also determine the content of assessment. When trying to diagnose potential learning handicaps, we still rely on aptitude-like measures designed to be as content-free as possible. We do so in order to avoid confusing lack of opportunity to learn with inability to learn. When the purpose of assessment is to measure actual learning, then content must naturally be tied to learning outcomes. However, even among achievement tests, there is considerable variability in the degree of alignment to a specific curriculum. Although to the lay person "math is math" and "reading is reading," measurement specialists are aware that tiny changes in test format can make a large difference in student performance. For example, a high proportion of students may be able to add numbers when they are presented in vertical format, but many will be unable to do the same problems presented horizontally. If manipulatives are used in some elementary classrooms but not in all, including the use of manipulatives in a mathematics assessment will disadvantage some children, while excluding their use will disadvantage others.

Assessments that are used to guide instruction in a given classroom should be integrally tied to the curriculum of that classroom. However, for large-scale assessments at the state and national level, the issues of curriculum match and the effect of assessment content on future instruction become much more problematic. For example, in a state with an agreed-upon curriculum, including geometry assessment in the early grades may be appropriate, but it would be problematic in states with strong local control of curriculum and so with much more curricular diversity.

Large-scale assessments, such as the National Assessment of Educational Progress, must include instructionally relevant content, but they must do so without conforming too closely to any single curriculum. In the past, this requirement has led to the problem of achievement tests that are limited to the "lowest common denominator." Should the instrument used for program evaluation include only the content that is common to all curricula? Or should it include everything that is in any program's goals? Although the common core approach can lead to a narrowing of curriculum when assessment results are associated with high stakes, including everything can be equally troublesome if it leads to superficial teaching in pursuit of too many different goals.

Finally, the intended use of an assessment will determine the need for normative information or other means to support the interpretation of assessment results. Identifying children with special needs requires normative data to distinguish serious physical, emotional, or learning problems from the wide range of normal development. When reporting to parents, teachers also need some idea of what constitutes grade-level performance, but such "norms" can be in the form of benchmark performances — evidence that children are working at grade level — rather than statistical percentiles.

To prevent the abuses of the past, the purposes and substance of early childhood assessments must be transformed. Assessments should be conducted only if they serve a beneficial purpose: to gain services for children with special needs, to inform instruction by building on what students already know, to improve programs, or to provide evidence nationally or in the states about programmatic needs. The form, substance, and technical features of assessment should be appropriate for the use intended for assessment data. Moreover, the methods of assessment must be compatible with the developmental level and experiences of young children. Below, I consider the implications of these principles for three different categories of assessment purposes.

Identifying Children with Special Needs

I discuss identification for special education first because this is the type of assessment that most resembles past uses of developmental screening measures. However, there is no need for wholesale administration of such tests to all incoming kindergartners. If we take the precepts

of developmentally appropriate practices seriously, then at each age level a very broad range of abilities and performance levels is to be expected and tolerated. If potential handicaps are understood to be relatively rare and extreme, then it is not necessary to screen all children for "hidden" disabilities. By definition, serious learning problems should be apparent. Although it is possible to miss hearing or vision problems (at least mild ones) without systematic screening, referral for evaluation of a possible learning handicap should occur only when parents or teachers notice that a child is not progressing normally in comparison to age-appropriate expectations. In-depth assessments should then be conducted to verify the severity of the problem and to rule out a variety of other explanations for poor performance.

For this type of assessment, developmental measures, including I.Q. tests, continue to be useful. Clinicians attempt to make normative evaluations using relatively curriculum-free tasks, but today they are more likely to acknowledge the fallibility of such efforts. For such difficult assessments, clinicians must have specialized training in both diagnostic assessment and child development.

When identifying children with special needs, evaluators should use two general strategies in order to avoid confounding the ability to learn with past opportunity to learn. First, as recommended by the National Academy Panel on Selection and Placement of Students in Programs for the Mentally Retarded,[11] a child's learning environment should be evaluated to rule out poor instruction as the possible cause of a child's lack of learning. Although seldom carried out in practice, this evaluation should include trying out other methods to support learning and possibly trying a different teacher before concluding that a child can't learn from ordinary classroom instruction. A second important strategy is to observe a child's functioning in multiple contexts. Often children who appear to be impaired in school function well at home or with peers. Observation outside of school is critical for children from diverse cultural backgrounds and for those whose home language is not English. The NAEYC stresses that "screening should never be used to identify second language learners as 'handicapped,' solely on the basis of their limited abilities in English."[12]

In-depth developmental assessments are needed to ensure that children with disabilities receive appropriate services. However, the diagnostic model of special education should not be generalized to a larger population of below-average learners, or the result will be the reinstitution of tracking. Elizabeth Graue and I analyzed recent efforts to create "at-risk" kindergartens and found that these practices are especially likely to occur when resources for extended-day programs are available only for the children most in need.[13] The result of such programs is often to segregate children from low socioeconomic backgrounds into classrooms where time is spent drilling on low-level prereading skills like those found on readiness tests. The consequences of dumbed-down instruction in kindergarten are just as pernicious as the effects of tracking at higher grade levels, especially when the at-risk kindergarten group is kept together for first grade. If resources for extended-day kindergarten are scarce, one alternative would be to group children heterogeneously for half the day and then, for the other half, to provide extra enrichment activities for children with limited literacy experiences.

Classroom Assessments

Unlike traditional readiness tests that are intended to predict learning, classroom assessments should support instruction by modeling the dimensions of learning. Although we must allow considerable latitude for children to construct their own understandings, teachers must nonetheless have knowledge of normal development if they are to support children's extensions and next steps. Ordinary classroom tasks can then be used to assess a child's progress in relation to a developmental continuum. An example of a developmental continuum would be that of emergent writing, beginning with scribbles, then moving on to pictures and random letters, and then proceeding to some letter/word correspondences. These continua are not rigid, however, and several dimensions running in parallel may be necessary to describe growth in a single content area. For example, a second dimension of early writing — a child's ability to invent increasingly elaborated stories when dictating to an adult — is not dependent on mastery of writing letters, just as listening comprehension, making predictions about books, and story retellings should be developed in parallel to, not after, mastery of letter sounds.

Although there is a rich research literature documenting patterns of emergent literacy and numeracy, corresponding assessment materials are not so readily available. In the next few years, national interest in developing alternative, performance-based measures should generate more materials and resources. Specifically, new Chapter 1 legislation is likely to support the development of reading assessments that are more authentic and instructionally relevant.

For example, classroom-embedded reading assessments were created from ordinary instructional materials by a group of third-grade teachers in conjunction with researchers at the Center for Research on Evaluation, Standards, and Student Testing.[14] The teachers elected to focus on fluency and making meaning as reading goals; running records and story summaries were selected as the methods of assessment.

But how should student progress be evaluated? In keeping with the idea of representing a continuum of proficiency, third-grade teachers took all the chapter books in their classrooms and sorted them into grade-level stacks, 1-1 (first grade, first semester), 1-2, 2-1, and so on up to fifth grade. Then they identified representative or marker books in each category to use for assessment. Once the books had been sorted by difficulty, it became possible to document that children were reading increasingly difficult texts with understanding. Photocopied pages from the marker books also helped parents see what teachers considered to be grade-level materials and provided them with concrete evidence of their child's progress. Given mandates for student-level reporting under Chapter 1, state departments of education or test publishers could help develop similar systems of this type with sufficient standardization to ensure comparability across districts.

In the meantime, classroom teachers — or preferably teams of teachers — are left to invent their own assessments for classroom use. In many schools, teachers are already working with portfolios and developing scoring criteria. The best procedure appears to be having grade-level

teams and then cross-grade teams meet to discuss expectations and evaluation criteria. These conversations will be more productive if, for each dimension to be assessed, teachers collect student work and use marker papers to illustrate continua of performance. Several papers might be used at each stage to reflect the tremendous variety in children's responses, even when following the same general progression.

Benchmark papers can also be an effective means of communicating with parents. For example, imagine using sample papers from grades K-3 to illustrate expectations regarding "invented spelling." Invented spelling or "temporary spelling" is the source of a great deal of parental dissatisfaction with reform curricula. Yet most parents who attack invented spelling have never been given a rationale for its use. That is, no one has explained it in such a way that the explanation builds on the parents' own willingness to allow successive approximations in their child's early language development. They have never been shown a connection between writing expectations and grade-level spelling lists or been informed about differences in rules for first drafts and final drafts. Sample papers could be selected to illustrate the increasing mastery of grade-appropriate words, while allowing for misspellings of advanced words on first drafts. Communicating criteria is helpful to parents, and, as we have seen in the literature on performance assessment, it also helps children to understand what is expected and to become better at assessing their own work.

Monitoring National and State Trends

In 1989, when the President and the nation's governors announced "readiness for school" as the first education goal, many early childhood experts feared the creation of a national test for school entry. Indeed, given the negative history of readiness testing, the first thing the Goal 1 Technical Planning Subgroup did was to issue caveats about what an early childhood assessment must *not* be. It should not be a one-dimensional, reductionist measure of a child's knowledge and abilities; it should not be called a measure of "readiness" as if some children were not ready to learn; and it should not be used to "label, stigmatize, or classify any individual child or group of children."[15]

However, with this fearsome idea set aside, the Technical Planning Subgroup endorsed the idea of an early childhood assessment system that would periodically gather data on the condition of young children as they enter school. The purpose of the assessment would be to inform public policy and especially to help "in charting progress toward achievement of the National Education Goals, and for informing the development, expansion, and/or modification of policies and programs that affect young children and their families."[16] Assuming that certain safeguards are built in, such data could be a powerful force in focusing national attention and resources on the needs of young children.

Beginning in 1998-99, a representative sample of 23,000 kindergarten students will be assessed and then followed through grade 5.

Unlike past testing practices aimed at evaluating individual children in comparison with normative expectations, a large-scale, nationally representative assessment would be used to monitor national trends. The purpose of such an assessment would be analogous to the use of the National Assessment of Educational Progress (NAEP) to measure major shifts in achievement patterns. For example, NAEP results have demonstrated gains in the achievement of black students in the South as a result of desegregation, and NAEP achievement measures showed gains during the 1980s in basic skills and declines in higher-order thinking skills and problem solving. Similar data are not now available for preschoolers or for children in the primary grades. If an early childhood assessment were conducted periodically, it would be possible to demonstrate the relationship between health services and early learning and to evaluate the impact of such programs as Head Start.

In keeping with the precept that methods of assessment should follow from the purpose of assessment, the Technical Planning Subgroup recommended that sampling of both children and assessment items be used to collect national data. Sampling would allow a broad assessment of a more multifaceted content domain and would preclude the misuse of individual scores to place or stigmatize individual children. A national early childhood assessment should also serve as a model of important content. As a means to shape public understanding of the full range of abilities and experiences that influence early learning and development, the Technical Planning Subgroup identified five dimensions to be assessed: 1) physical well-being and motor development, 2) social and emotional development, 3) approaches toward learning, 4) language usage, and 5) cognition and general knowledge.

Responding to the need for national data to document the condition of children as they enter school and to measure progress on Goal 1, the U.S. Department of Education has commissioned the Early Childhood Longitudinal Study: Kindergarten Cohort. Beginning in the 1998-99 school year, a representative sample of 23,000 kindergarten students will be assessed and then followed through grade 5. The content of the assessments used will correspond closely to the dimensions recommended by the Technical Planning Subgroup. In addition, data will be collected on each child's family, community, and school/program. Large-scale studies of this type serve both program evaluation purposes (How effective are preschool services for children?) and research purposes (What is the relationship between children's kindergarten experiences and their academic success throughout elementary school?).

National needs for early childhood data and local needs for program evaluation information are similar in some respects and dissimilar in others. Both uses require group data. However, a critical distinction that affects the methods of evaluation is whether or not local programs share a

Fearing that "assessment" is just a euphemism for more bad testing, many early childhood professionals have asked, Why test at all?

common curriculum. If local programs, such as all the kindergartens in a school district, have agreed on the same curriculum, it is possible to build program evaluation assessments from an aggregation of the measures used for classroom purposes. Note that the entire state of Kentucky is attempting to develop such a system by scoring classroom portfolios for state reporting.

If programs being evaluated do not have the same specific curricula, as is the case with a national assessment and with some state assessments, then the assessment measures must reflect broad, agreed-upon goals without privileging any specific curriculum. This is a tall order, more easily said than done. For this reason, the Technical Planning Subgroup recommended that validity studies be built into the procedures for data collection. For example, pilot studies should verify that what children can do in one-on-one assessment settings is consistent with what they can do in their classrooms, and assessment methods should always allow children more than one way to show what they know.

Conclusion

In the past decade, testing of 4-, 5-, and 6-year-olds has been excessive and inappropriate. Under a variety of different names, leftover I.Q. tests have been used to track children into ineffective programs or to deny them school entry. Prereading tests held over from the 1930s have encouraged the teaching of decontextualized skills. In response, fearing that "assessment" is just a euphemism for more bad testing, many early childhood professionals have asked, Why test at all? Indeed, given a history of misuse, the burden of proof must rest with assessment advocates to demonstrate the usefulness of assessment and to ensure that abuses will not recur. Key principles that support responsible use of assessment information follow.

- No testing of young children should occur unless it can be shown to lead to beneficial results.
- Methods of assessment, especially the language used, must be appropriate to the development and experiences of young children.
- Features of assessment — content, form, evidence of validity, and standards for interpretation — must be tailored to the specific purpose of an assessment.
- Identifying children for special education is a legitimate purpose for assessment and still requires the use of curriculum-free, aptitude-like measures and normative comparisons. However, handicapping conditions are rare; the diagnostic model used by special education should not be generalized to a larger population of below-average learners.
- For both classroom instructional purposes and purposes of public policy making, the content of assessments should embody the important dimensions of early learning and development. The tasks and skills children are asked to perform should reflect and model progress toward important learning goals.

In the past, local newspapers have published readiness checklists that suggested that children should stay home from kindergarten if they couldn't cut with scissors. In the future, national and local assessments should demonstrate the richness of what children do know and should foster instruction that builds on their strengths. Telling a story in conjunction with scribbles is a meaningful stage in literacy development. Reading a story in English and retelling it in Spanish is evidence of reading comprehension. Evidence of important learning in beginning mathematics should not be counting to 100 instead of to 10. It should be extending patterns; solving arithmetic problems with blocks and explaining how you got your answer; constructing graphs to show how many children come to school by bus, by walking, by car; and demonstrating understanding of patterns and quantities in a variety of ways.

In classrooms, we need new forms of assessment so that teachers can support children's physical, social, and cognitive development. And at the level of public policy, we need new forms of assessment so that programs will be judged on the basis of worthwhile educational goals.

1. Lorrie A. Shepard and Mary Lee Smith, "Escalating Academic Demand in Kindergarten: Counterproductive Policies," *Elementary School Journal*, vol. 89, 1988, pp. 135-45.
2. Sue Bredekamp, ed., *Developmentally Appropriate Practice in Early Childhood Programs Serving Children from Birth Through Age 8*, exp. ed. (Washington, D.C.: National Association for the Education of Young Children, 1987).
3. M. Therese Gnezda and Rosemary Bolig, *A National Survey of Public School Testing of Pre-Kindergarten and Kindergarten Children* (Washington, D.C.: National Forum on the Future of Children and Families, National Research Council, 1988).
4. Samuel J. Meisels, "Uses and Abuses of Developmental Screening and School Readiness Testing," *Young Children*, vol. 42, 1987, pp. 4-6, 68-73.
5. Lorrie A. Shepard and M. Elizabeth Graue, "The Morass of School Readiness Screening: Research on Test Use and Test Validity," in Bernard Spodek, ed., *Handbook of Research on the Education of Young Children* (New York: Macmillan, 1993), pp. 293-305.
6. Anne C. Stallman and P. David Pearson, "Formal Measures of Early Literacy," in Lesley Mandel Morrow and Jeffrey K. Smith, eds., *Assessment for Instruction in Early Literacy* (Englewood Cliffs, N.J.: Prentice-Hall, 1990), pp. 7-44.
7. Lorrie A. Shepard, "A Review of Research on Kindergarten Retention," in Lorrie A. Shepard and Mary Lee Smith, eds., *Flunking Grades: Research and Policies on Retention* (London: Falmer Press, 1989), pp. 64-78.
8. Lorrie A. Shepard and Mary Lee Smith, "Academic and Emotional Effects of Kindergarten Retention in One School District," in idem, pp. 79-107.
9. "Guidelines for Appropriate Curriculum Content and Assessment in Programs Serving Children Ages 3 Through 8," *Young Children*, vol. 46, 1991, pp. 21-38.
10. Ibid., p. 32.
11. Kirby A. Heller, Wayne H. Holtzman, and Samuel Messick, eds., *Placing Children in Special Education* (Washington, D.C.: National Academy Press, 1982).
12. "Guidelines," p. 33.
13. Shepard and Graue, op. cit.
14. The Center for Research on Evaluation, Standards, and Student Testing is located on the campuses of the University of California, Los Angeles, and the University of Colorado, Boulder.
15. *Goal 1: Technical Planning Subgroup Report on School Readiness* (Washington, D.C.: National Education Goals Panel, September 1991).
16. Ibid., p. 6. ■