BTJ Editor's note: This Public Policy Viewpoint focuses on issues related to the assessment of young children and the new assessment requirements for children in Head Start programs. At the time of this Web posting (1/13/04), Head Start programs have undergone a significant event: use of a federally created test to assess the language, literacy, and math knowledge of all four-year-old children enrolled in Head Start. The Head Start Child Outcomes Framework, Program Performance Standards, and teacher qualifications requirements have influenced child care and state prekindergarten programs. As the Head Start reauthorization and this new assessment move forward, it is important to consider how testing young children may affect the future of early childhood education. We welcome your views on this topic.

The Head Start National Reporting System
A Critique

Samuel J. Meisels and Sally Atkins-Burnett

Initiated in the fall of 2003, a high-stakes achievement test is being administered to all four- and five-year-olds in Head Start. Two times a year, 35,000 teachers or surrogates will administer the 15- to 20-minute test to more than a half-million children, at a cost in excess of $16 million. The purpose of the test, as described by the Head Start Bureau, is threefold: (1) to enhance local aggregation of child outcome data and local program self-assessment efforts; (2) to enable the Head Start Bureau and Administration for Children and Families (ACF) Regional Offices to plan training and technical assistance efforts; and (3) to incorporate child outcome information into future Head Start program monitoring reviews. Never before in the history of this nation have so many young children been exposed to a standardized achievement test.

Samuel J. Meisels, Ed.D., is president of Erikson Institute, a graduate school in child development located in Chicago. A former faculty member and currently an emeritus professor at the University of Michigan, he is one of the nation's leading authorities on the assessment of young children. He is coauthor of the *Handbook of Early Childhood Intervention, The Work Sampling System, The Ounce Scale,* and many other publications.

Sally Atkins-Burnett, Ph.D., is an assistant professor of early childhood and special education at the University of Toledo. She was involved in the development of the child assessments for the Early Childhood Longitudinal Study—Kindergarten Cohort and provides consultation to other national longitudinal studies regarding child assessments. She teaches graduate level courses in early childhood assessment.

Unfortunately, this test, called the National Reporting System (NRS), includes items that are rife with class prejudice and are developmentally inappropriate. This is particularly troubling because the test is used by Head Start officials as a quality assurance system. In fact, the idea that a narrow test of young children's skills in literacy and math can represent a quality indicator of a holistic program like Head Start shows a stunning lack of appreciation for the comprehensive goals of the 38-year-old program. Moreover, program quality cannot be evaluated by student outcomes alone, since this approach does not take into account differences among children and programs.

A problematic approach

The approach to program evaluation that the Head Start Bureau has chosen is extremely problematic. Researchers have demonstrated repeatedly that testing children using the design of the NRS—an isolated evaluation of children's skills at two time points—is highly inaccurate and renders very poor predictions of later achievement (LaParo & Pianta 2000). Young children's skills are in flux; this lack of stability renders the conclusions based on con-

ventional pre/post tests potentially misleading. Moreover, the Head Start population, composed primarily of children whose families have incomes below the national poverty line (approximately $18,500 a year for a family of four in 2003), is very diverse in terms of ethnicity, languages spoken at home, parental education, and concomitant opportunities for children to learn prior to entering Head Start. Both the National Education Goals Panel (Shepard, Kagan, & Wurtz 1998) and the National Research Council of the National Academies (Heubert & Hauser 1999) urge that children below age eight not be administered the kinds of tests that are represented by the NRS. Even the No Child Left Behind Act does not mandate testing until third grade. For many negative reasons, the NRS breaks new ground.

Description of the test

The NRS consists of five subtests: two measure English-language competence, one evaluates receptive vocabulary knowledge, one focuses on letter names, and the final subtest addresses mathematics.

The test begins with two subtests that are intended to determine if a child has sufficient mastery of the English language to take the rest of the test. If the child does not demonstrate this mastery, and is a Spanish speaker, a Spanish-language version of the test is administered. However, before the decision is made to administer the test in Spanish, the child must miss more than 70 percent of the items on these two subtests. This means the child experiences at least 15 failures. If the tester knows that the child comes from a Spanish-speaking family, but the child passes the language screen, the child is administered both language versions. If the child speaks any of the roughly one hundred other languages represented in Head Start, the child is not tested, because the test is available only in Spanish and English.

The vocabulary subtest is adapted from the Peabody Picture Vocabulary Test-III (PPVT-III), a test of receptive language skills such as listening comprehension. Children are shown a page with four pictures; the examiner says a word and asks the child to point to the picture that "best shows what the word means." The NRS version contains 24 items. Although the measure includes items that appropriately ask children to identify body parts, animals, and actions, some items are less appropriate, including a number that are very class-biased. For example, for the word *vase* the incorrect illustrations, or test foils, include items that could actually all be used as vases.

An item that requires children to select the facial expression for *horrified* from an array of four faces is highly inappropriate for four-year-olds. The facial features of all four images are Caucasian, ignoring the fact that facial expressions differ in different cultural and ethnic groups. Also, by including this emotion, the test ignores extensive research indicating that the depiction of the facial expression for *horrified* (which researchers note is often confused with the expression of anger or rage) is not recognized by most children until much later in life (Izard 1977, 1991; DePaulo & Rosenthal 1982). This is the only emotional expression that children are asked to identify in the assessment.

When the entire PPVT-III is administered, differences in the general knowledge and experience of children taking the test are balanced by the broad range of vocabulary items on the test. In a subset of about a third of the items for the age ranges selected for the NRS (2½ to 9), such differences are highlighted; and these items clearly place children from impoverished backgrounds at a significant disadvantage. Further, the items are administered in an order that differs from the way they are presented on the original PPVT-III, which may result in a further loss of validity for this group of items. The evidence from the field test indicated lower reliability for this subset of items (below .80 for interrater reliability, or consistency, and for the Item Response Theory [IRT]-based statistical reliability when examined by age) than for the test as a whole.

The letter-naming task on the vocabulary subtest is totally removed from children's natural use of and experience with letters. Except on a bulletin board in an elementary school, children seldom see an alphabet with both upper- and lowercase letters displayed in pairs. Further, children are rarely exposed to letters arrayed in matrixes that contain nine or more of these pairs. Children see letters embedded in words and connected to actual objects, signs, or books. The "principle of contextualization" (Messick 1983) reminds us that the meaning of a test item is changed when that item is taken out of context or placed in a new context. No appreciation of context is recognized on this subtest.

The principal reason for assessing children's ability to name letters, other than learning about children's memory skills and exposure to letters, is to determine how quickly a child can name letters. Rapid letter naming is a marker of a child's ability to encode symbols, is related to phonemic awareness, and is predictive of later reading ability (Adams 1990; Lombardino et al. 1999; Uhry 2002). But the NRS is not a power test and this subtest does *not* measure how quickly children can complete the task. In fact, the directions suggest to children that they not rush to respond to the various displays of letters. In addition, the child receives credit only for naming the letter, not for providing the sound the letter represents. In short, because of the way this test is administered, it may inform us of nothing more than the extent of a child's opportunity to

learn—and provides limited information even about that—while ignoring research about the importance of phoneme and sound awareness.

The math subtest is the most problematic, as it is more a test of language competence than quantitative skills. Approximately 25 percent of this subtest involves naming a number or pointing to a shape that is named. It also contains several if-then statements and terms like *longer than* and other comparatives. Further, the items are very poorly designed. For example, children are asked to count grapes, but some of the grapes are shown in a cluster, some individually. The point is to assess mathematical skills, not general knowledge.

The test does not provide a correct answer to one of the questions. It asks which crayon from an array of four crayons is longer than a brush, but the *brush* that is shown is shorter than all the crayons. If the item had asked for a comparison of the crayons with the *paintbrush* this question could be answered unambiguously, but this is not what the item asks for. It requires a child to know that the word *brush* refers to the handle *and* brush as a single item. Children who are unfamiliar with English shorthand may be confused by the question and not be able to answer correctly. The item could be improved by using a pencil or something else with only one part.

Another anomalous item shows several coins and asks, "Which coin is smaller in size than the penny?" This question is extremely difficult because it requires that children make a distinction between the physical size of a coin and its monetary value, be familiar with five different coins (one of which is the relatively rare 50-cent piece), and know the meaning of *which*.

In general, the assessment of number concepts is very limited. Only two items require a child to have a concept of any number greater than three. There are no items that assess comprehension of pattern, number constancy, matching, classification, or estimation; spatial reasoning; or recognition of more and less, all of which are central to the development of young children's mathematical thinking.

The if-then construction used in three questions (for example, "If you gave a friend one of these books, how many would you have left?") is difficult for children to understand before elementary school and is particularly language driven. Most children are at least school age before they are competent in understanding sentences that are joined with such conjunctions as *if, because,* and *unless* (Gummersall & Strong 1999; Hulit & Howard 2002). The complexity of the sentence construction distracts from the assessment of mathematical thinking. The test's measurement item that refers to inches is also not considered appropriate before second or third grade, according to the National Council of Teachers of Mathematics (2000) and other researchers (Kribs-Zaleta & Bradshaw 2003). The final item, which requires mastery of one-to-one correspondence while counting, is so difficult to score that reliability among examiners is likely to be very low.

Conclusion

In short, this test teaches us very little about young children's preschool skills. It provides no authentic literacy evaluation and little information about math skills. Entire areas of development, such as social-emotional growth, physical development, science, social studies, the arts, and most of literacy and even phonemic awareness, are omitted. Basic concepts are included only in the initial language mastery items, and not only are these very limited, it is not clear that this subtest will be reported in terms of anything other than English-language mastery.

The vocabulary test, which contains many items that are culturally specific, may have secondary negative effects. Because so many Head Start teachers have restricted or narrow training (fewer than 30 percent hold a B.A.), they may consider the items on this test to be appropriate for all children to know and this may greatly distort their curriculum. As one trainer in a Head Start agency said, "This is the first time the Head Start Bureau has directed us to do something that can really do damage to children."

The NRS is biased toward children from families with high socioeconomic status and Caucasian culture. Not only is it class- and race-oriented, it also reflects a singular pedagogical approach. It has few expressive and no creative, constructive, or active elements; it encourages a passive transmission model of education. It offers no problem-solving items, does not call for complex or higher-order thinking, and allows for few if any alternative answers. The lessons it teaches Head Start staff are negative regarding children's potential, and it does not recognize the richness of the backgrounds that Head Start children bring with them. Any conclusions regarding the quality of Head Start that are based on data from this test will be tenuous at best.

Because the NRS will be used in making decisions about the continuation of Head Start, it is an example of a test in which indicators of learning can overwhelm learning itself, since potentially successful programs may be classified as failures based on an invalid and poorly constructed test. After all the experience this nation has had regarding the impact of measurement-driven instruction in the K–12 arena, it is baffling that a test like this would be visited on young children from poor households.

This test is not good early education practice. It is not good psychometric practice. It is not good public policy. And it is certainly not good for young children.

References

Adams, M.J. 1990. *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

DePaulo, B.M., & R. Rosenthal. 1982. Measuring the development of sensitivity to

nonverbal communication. In *Measuring emotions in infants and children,* ed. C.E. Izard, 208–47. New York: Cambridge University Press.
Gummersall, D.M., & C.J. Strong. 1999. Assessment of complex sentence production in a narrative context. *Language, speech, and hearing services in Schools* 30 (2): 152–64.
Heubert, J.P., & R.M. Hauser, eds. 1999. *High stakes: Testing for tracking, promotion, and graduation.* Board on Testing and Assessment, National Research Council. Washington, DC: National Academy Press.
Hulit, L.M., & M.R. Howard. 2002. *Born to talk: An introduction to speech and language development.* 3d ed. Boston: Allyn & Bacon.
Izard, C.E. 1977. *Human emotions.* New York: Plenum.
Izard, C.E. 1991. *The psychology of emotions.* New York: Plenum.
Kribs-Zaleta, C.M., & D. Bradshaw. 2003. A case of units. *Teaching Children Mathematics* 9 (7): 397–99.
LaParo, K.M., & R.C. Pianta. 2000. Predicting children's competence in the early school years. A meta-analytic review. *Review of Educational Research* 70 (4): 443–84.
Lombardino, L.J., D. Morris, L. Mercado, F. DeFillipo, C. Sarisky, & A. Montgomery. 1999. The Early Reading Screening Instrument: A method for identifying kindergartners at risk for learning to read. *International Journal of Language and Communication Disorders* 34 (2): 135–50.
Messick, S. 1983. Assessment of children. In *Handbook of child psychology: History, theory, and methods,* Vol. 1, ed. W. Kessen, 477–526. New York: Wiley.
NCTM (National Council of Teachers of Mathematics). 2000. *Principles and standards for school mathematics.* Reston, VA: Author.
Shepard, L.A., S.L. Kagan, & E. Wurtz, eds. 1998. *Principles and recommendations for early childhood assessments.* Prepared for the National Education Goals Panel by the Goal 1 Early Childhood Assessments Resource Group. Washington, DC: Government Printing Office.
Uhry, J.K. 2002. Finger-point reading in kindergarten: The role of phonemic awareness, one-to-one correspondence, and rapid serial naming. *Scientific Studies of Reading* 6 (4): 319–42.