The Head Start National Reporting System As A Model for Systems Aimed at Assessing and Monitoring the Performance of Preschool Programs

Nicholas Zill, Ph.D.
Child and Family Studies
Westat

Paper prepared for the Pew Foundation National Early Childhood Accountability Task Force
November 13, 2007

Introduction

In recent years, a growing number of states in the U.S. have chosen to provide publicly-funded preschool programs for their pre-kindergarten-aged children. Along with the expanded availability of public preschool have come increased calls for these programs to be held accountable for achieving measurable results in bolstering the skills and school readiness of the children who attend them. But how should the contribution of these programs to child development be assessed and evaluated? The Head Start National Reporting System (NRS) is one response to these demands for greater accountability, one focused on the largest federal preschool program for children from low-income families. The NRS was designed to provide indicators of the progress children are making on key early literacy and math skills for all local Head Start programs.

One of the major reasons for having 3-, 4- and 5-year-old children participate in preschool programs is to have them acquire early skills and knowledge that will help them succeed in elementary school and beyond, skills that they might not acquire if they remained at home or participated in child care programs without an educational component. Therefore, in judging how well a given preschool program is doing its job, it is useful to know what skills and knowledge children have, on average, when they enter the program and what skills they have when they complete the program and go on to kindergarten.

Whether or not they attend preschool, young children can be expected to show growth over time in many skill areas. Thus, in order to evaluate the extent to which a local program bolsters growth, it is desirable to be able to compare the skills shown by children who participate in the program with the skills shown by children in one or more reference groups. The most desirable reference group, but one that is difficult to obtain, is a sample of children who are similar to those in the program but who have been randomly selected not to participate in the preschool program in question (a random-assignment control group). This kind of comparison is available, on an aggregate, nationwide level, in the National Head Start Impact Study (Administration for Children and Families, 2005). But creating such comparison groups for every local Head Start program was obviously not practical. Local Head Start programs can be compared with one another, however. And children who attended a local Head Start program in one year can be compared with those who attended the same program in earlier years.

In order to make these potential comparisons, the skill and knowledge levels of children in the local Head Start program and children in the reference groups must be determined in the same way, so that the achievements of the different groups may be legitimately compared. The NRS was intended to facilitate such comparisons by taking steps to standardize the aspects of child development that are assessed, the conditions of assessment, and the rules for scoring children's responses. This served to maximize comparability between assessment results of different children done by different people (assessors) at different times and places. The NRS provided Head Start, for the first time in its forty-year history, with consistently-collected, comparable achievement measures for all local programs.

The NRS was not designed to help guide instruction, diagnose disabilities, or guide the placement of individual children. Nor was it intended as some sort of kindergarten readiness test. Rather, its purpose was to provide aggregate program-level indicators of the levels of achievement of groups of children when they entered and left specific Head Start programs, and indicators of the growth in knowledge they showed from early in the program year to the end of that year. It was designed to cover a limited but important set of knowledge areas. The results of the assessments were to be used in planning training and technical assistance efforts for local programs, so that they might do a better job of bolstering children's achievement.

Direct Assessment versus Observation-Based Ratings

Before embarking on the design of a National Reporting System based on direct assessment of children's knowledge and skills, the Office of Head Start surveyed the child observation and assessment practices being used by local Head Start programs throughout the country. This was done to find out whether it might be feasible to combine local assessment results in such a way as to make them comparable from program to program. A second purpose was to determine whether it seemed practical to develop a reporting system based on observation of children's skills in the classroom rather than on direct assessment of those skills. The conclusion of the survey was that neither of these alternatives seemed likely to be as practical or cost-efficient as a system using standardized direct assessments of children's knowledge and skills.

The great advantage of using direct child assessment to measure children's skill levels is that steps can be taken to standardize the aspects of child development that are assessed, the conditions of assessment, and the rules for scoring children's responses. This serves to maximize comparability between assessment results done by different people (assessors) at different times and places.

Standardizing the conditions of an observation is much more difficult to accomplish. Different preschool curricula differ in the kinds of child instructional activities they emphasize. Thus, they may also differ in the opportunities they provide for a rater to observe a child demonstrating a given kind of skill. This suggests that it would be more difficult to make outcome comparisons across preschool programs with different curricula using observational methods than by using direct assessment methods.

Furthermore, with direct assessment, there are well worked-out procedures for determining the reliability of the assessment process, both in terms of the internal consistency of the results and the extent of agreement of assessments done by different assessors or by the same assessor on different occasions. These procedures can be applied in a relatively straightforward fashion, without requiring a great deal of additional time and effort.

By contrast, establishing the reliability of observational methods is often quite challenging, because it requires that two independent raters observe a given child on a number of different occasions over a relatively extended period of time, while the child is engaged in the same kinds of ongoing classroom activities. This can certainly be done, but it is more difficult to arrange and winds up being considerably more intrusive and burdensome for a preschool program than comparable procedures for establishing the reliability of direct assessment.

Differences between direct assessment and observation approaches are also observed with respect to training and certifying staff members to carry out these respective methods with accuracy and reliability. It is generally quicker and easier to train a person to do direct child assessment reliably than to do observation-based ratings with adequate reliability. And procedures for certifying that a trainee has learned the method satisfactorily are also simpler and faster to accomplish with direct assessment than with naturalistic observation. This is especially true if the direct assessment procedures have been simplified so as not to require a lot of judgment calls and branching decisions on the part of the assessor-trainee.

How the National Reporting System Works

In the NRS, all children in participating Head Start programs who are 4- or 5-years-old and are eligible to attend kindergarten in the following academic year are given the same brief, one-on-one assessment. The assessment is carried out by a trained adult assessor from the local program who interacts with the child for about 20 minutes in a room or cubicle that is reasonably free from distractions. During this time, the assessor shows the child a series of plates and asks the child questions about them. The skills assessed involve such tasks as understanding and following spoken instructions, naming common objects, telling which of several pictures best shows the meaning of a given word, recognizing letters of the alphabet, doing one-to-one counting, recognizing geometric shapes, making relative size judgments, and solving simple addition and subtraction problems.

The child's responses are recorded by the assessor on a scannable answer sheet. The answer sheets for all assessed children in the program are assembled and sent to a central processing facility, where they are scanned. The resulting response data are electronically transmitted to an NRS analysis and reporting center. There the response data are scored using Item Response Theory-based scoring programs. Scores for all assessed children in each program are aggregated and summary statistics about the average performance and range of skills shown by children in the local program are transmitted back to the program, as well as to regional and national bureaus of the Office of Head Start.

Similar assessments are given to the children in fall and spring and the results are reported to the programs to show the progress of their children over the course of the year. The tasks presented to the children differ somewhat from fall to spring to try to forestall narrow "teaching to the test." But the items making up the assessments are selected so that their average difficulty level is equivalent from fall to spring. Likewise, while the item composition of NRS assessments varies somewhat from year to year, efforts are made to ensure that average difficulty levels and the range of item difficulties remain comparable. Thus, Head Start programs can compare the achievement levels and growth of their children from one cohort to the next, as well as comparing the progress of children in their program with that of children in other local programs.

Computer-Based Reporting System. The NRS makes use of a Computer-Based Reporting System (CBRS) to register and track children who are eligible to be assessed and program staff members who are eligible to act as assessors. Local program staff members enter information via the Internet on the number of children in the program eligible for assessment, basic demographic information about each child, and information about the center and class which they are attending. Local administrative staff also enter information about the child's lead teacher and the person who will be assessing the child. If the local program already has a computer-based program information system, data from that system can usually be imported directly into the CBRS, without the need for duplicative data re-entry. Programs that lack adequate computer or Internet capabilities can enter the necessary information on paper forms and send those forms to the NRS analysis center for conversion to electronic records.

As assessments are completed, local administrative staff enter information indicating whether each child was assessed and, if not, the reason for non-assessment. Demographic information about the program, center, class, teacher, child and assessor are combined with scanned assessment records to form an integrated data base. The CBRS is also used to transmit preliminary reports on NRS assessment results to local programs in a timely manner. Once final versions of program reports are completed, they are transmitted to the programs both electronically via the CBRS and as hard-copy reports sent by regular mail.

Data encryption. In order to protect the privacy of children and adults involved in the NRS and prevent unauthorized use of collected information, assessment and demographic data are encrypted and unique identification numbers (rather than names) are used to label individual children, teachers, and assessors. Only the local Head Start programs have the ability to link names of individual pupils or staff members with NRS identification numbers. The NRS analysis center and the Office of Head Start are not able to link names with numbers.

Training and certification of assessors. The NRS makes use of a "train the trainers" approach to prepare assessors from local Head Start programs to be able to conduct child assessments in a standardized manner. During the summer preceding the first year of national implementation of the NRS, eight regional training sessions were held to train one or more NRS coordinators and lead trainers from some 2,000 local programs. The training sessions were three days long and involved classroom lectures, reading training manuals, watching videos of properly-conducted assessments, role plays, and practice assessments of actual children. Classes were also conducted on procedures for entering data and receiving reports via the CBRS.

Each trainee had to pass a paper-and-pencil test on assessment procedures as well as being observed and scored for accuracy and rapport while she or he conducted an assessment of a child. The trainee had to achieve a score of 85 or better (out of 100) on both the written test and the child assessment in order to be certified. Bilingual assessment trainees had to be certified in both Spanish- and English-language child assessments. Trainees who passed were given practice and feedback as they conducted mock training sessions of their own. They then returned to their local programs and trained other local staff members to conduct child assessments in appropriate fashion. In all, more than 20,000 local assessors were trained and certified.

During subsequent summers, similar but smaller-scale regional trainings have been held to train individuals to serve as replacement NRS lead trainers for local programs who have lost their original lead trainers through staff turnover. In addition, preceding each round of the NRS assessments, local programs are sent updates to the assessment and CBRS training manuals and videos that demonstrate the appropriate administration of new or modified assessment items. The refresher videos also provide guidance on dealing with child problem behavior during assessments and other issues.

Selection of assessors. The selection of staff to serve as NRS assessors has been left up to local program administrators. They could make use of the child's own lead teacher as an assessor, another teacher, another staff member, such as an education coordinator, or even an outside consultant, as they see fit. In Year Three of the NRS (2005-2006), 75 percent of the assessments were conducted by someone other than the child's own teacher. Whatever local administrators chose to do, they were requested to record the assessor's relationship to the assessed child as one of the required data elements in the CBRS. Of the 24,561 trained assessors registered in the CBRS in Year Three, 50 percent were Head Start teachers, 29 percent were other Head Start professional staff, 18 percent were other Head Start staff members, and 3 percent were consultants or graduate students. The availability of information about assessor identity meant that one could examine relationships that might emerge between the type of adult who conducts the assessment and means, variances, and reliabilities of children's assessment scores.

Numbers of programs and children involved. The scale of the NRS is unprecedented in terms of the numbers of preschool children involved. In the spring of 2006, for example, there were a total of 1,905 Head Start grantees or delegate agencies that had 4- or 5-year-old children enrolled and were eligible to participate in the NRS. Of these, 1,787, or 94 percent, did cooperate with the assessment program. In these programs, there were 406,263 children who were entered into the CBRS as eligible to be assessed. The number of these children who were assessed was 396,823, or 97.7 percent. There were 298,972 assessments conducted in English only, 22,903 conducted in Spanish only, and 74,948 children who were assessed in both English and Spanish. Some 71 percent of the children were from racial or ethnic minority groups, and 26 percent were English-language learners (ELLs). Nine percent of the children had identified disabilities.

Assessment Components and How They Were Selected

The NRS Assessment in English consisted of four components:

- ► a language-proficiency screener, to determine whether the child had sufficient understanding of spoken English for him or her to be assessed fairly in the remaining sections of the assessment;

- ► a vocabulary test, measuring the child's knowledge of the meaning of common English words;

- ► a letter naming test, measuring the child's ability to recognize and name all 26 letters of the English alphabet; and,

- ► an early math skills test, measuring the child's ability to read numerals, recognize geometric shapes, make relative size judgments, interpret graphical representations of quantities, and solve simple counting, addition, and subtraction problems.

Children from Spanish-speaking families were also administered a Spanish version of the NRS, which had four parallel components:

- ► a Spanish language-proficiency screener;

- ► a Spanish vocabulary test;

- ► a Spanish letter naming test; and

- ► an early math skills test in Spanish.

Children from English-speaking families were administered all four components of the English assessment, whether they passed the language screener or not. Children from Spanish-speaking families were administered the English-language screener. If they passed this, they were administered the other three components of the English assessment. Whether or not they passed the English screener, they were also administered the Spanish-language screener. If they passed this, they were administered the other three components of the Spanish assessment.

In the first year of national implementation of the NRS, children from Spanish-speaking families were administered the English-language screener first. After the first year of national implementation, the Office of Head Start made a policy decision to administer the Spanish-language screener and the remainder of the Spanish assessment to children from Spanish-speaking families before they were administered the English-language screener and assessment. Children in Head Start programs in Puerto Rico, where Spanish is the language of instruction, were administered the Spanish assessment only.

Children from non-English-speaking, non-Spanish-speaking families were administered the English-language screener. If they passed this, they were administered the other three components of the English assessment. If they did not pass the screener, the assessment was ended.

The procedures just described were followed in both fall and spring assessments. Thus, it would be possible for a child from a non-English-speaking family not to pass the English assessment in the fall, but to pass it in the spring, after he or she had had more exposure to spoken English. In that case, the child would have test scores on the three other English components (vocabulary, letter naming, early math) for the spring only.

Although the English NRS Assessment and the Spanish NRS Assessment had similar components, similar content areas, and similar items, no attempt was made to ensure that the two versions of the test had exactly equal difficulty levels. It would be very difficult to establish such equivalence, as it would require giving both English and Spanish versions of items in the test item pools to a large bilingual subject population composed of children who were equally proficient in Spanish and English. For reasons of practicality and cost, this was not done. It is reasonable to suppose, however, that because of their similar construction, the English and Spanish Assessments were of roughly equal difficulty.

The child assessment of the HSNRS was designed to provide direct measures of how well Head Start children had mastered some of the skills represented in the Head Start standards of learning, the Child Outcomes Framework (ACYF, 2003). However, the Office of Head Start recognized from the outset that there was no way that the NRS could cover all or even most of the many skills identified in the Outcomes Framework. The number and scope of measures had to be limited, given the need to develop a system that presented minimal burden to Head Start programs, staff, and children. Thus, the HSNRS was not developed to represent ALL skills that Head Start children should learn, but rather a few key indicators of children's language and literacy development.

Criteria for task selection. The selection of tasks to be included in the child assessment of the HSNRS was guided by the following criteria. Tasks in the assessment were intended to appraise skills that:

► Were critical stepping-stones on the path to achievement in elementary school, especially in the areas of reading and mathematics;

► Could be readily enhanced among preschoolers by activities in Head Start;

► Head Start parents wanted their children to learn;

► Congress expected children to learn in Head Start, as indicated by their being among the mandated achievement goals contained in legislation that reauthorized the Head Start program in 1998;

► A majority of U.S. children from non-low-income families had mastered by the time they begin Kindergarten; and,

► Could be reliably measured in a relatively brief child assessment that can be conducted by a Head Start teacher or other local staff member.

The NRS assessment battery was intended to measure key indicators of children's early language, literacy, and math skills development. At the same time, it had to be an instrument that: 1) was simple to administer; 2) collected valid and reliable data; 3) was at an appropriate difficulty level for Head Start children 4) included items that discriminate well between children at different levels of development; and 5) minimized the burden on local program staff in terms of training on and administration of the data collection procedures.

In choosing to focus on the content areas of reading and mathematics, the Office of Head Start benefited from the extensive base of theory that has been created and tested in these areas by psychologists and other educational researchers. It also benefited from prior longitudinal research studies on the development of reading and math skills from the preschool through the elementary and secondary school years, including studies like those of Whitehurst and his colleagues (Storch & Whitehurst, 2002) and the Head Start Family and Child Experiences Survey (FACES) (Administration for Children and Families, 2001, 2003; Zill & Resnick, 2006) and the National Head Start Impact Study (ACF, 2005), that focused specifically on Head Start children.

<u>Selection of domains of early achievement.</u> Based on the selection criteria described above, and making use of the measurement experience accumulated in the longitudinal studies just enumerated, the Office of Head Start selected the achievement domains of vocabulary, letter naming, early math skills, phonological awareness, and oral language proficiency for field testing and possible inclusion in the NRS.

► Vocabulary

Vocabulary is a central knowledge area which becomes particularly important when children make the transition from "learning to read" to "reading to learn" (Chall, 1983; Biemiller, 2006, p. 41; Whitehurst & Lonigan, 1998). Vocabulary measures are included in nearly all batteries of general cognitive functioning, intelligence, and language development, such as the Wechsler, Stanford-Binet, McCarthy, Kaufman, and Woodcock-Johnson Scales (Strauss, Sherman, & Spreen, 2006, pp. 98 - 400). In the 1998 legislation that reauthorized Head Start, the U.S. Congress mandated that Head Start ensure that each child, "Understands increasingly complex vocabulary."

► Letter Naming

The ability to identify the letters of the alphabet by name is one of the skills that are essential stepping-stones on the path to becoming a skilled reader (Whitehurst & Lonigan, 1998; Ehri & Roberts, 2006). Letter naming is a component of several widely-used measures of reading proficiency, such as the Letter-Word Identification subtest of the Woodcock-Johnson III Tests of Achievement, the Basic Reading subtest of the Wechsler Individual Achievement Test – II, and Reading subtest of the Wide Range Achievement Test – 3 (Strauss, Sherman, & Spreen, 2006, pp. 370-400). In the 1998 legislation that reauthorized Head Start, the U.S. Congress mandated that Head Start ensure that each child be able to identify at least 10 letters of the alphabet.

► Early Math

The early math skills of preschoolers are stepping stones toward and predictive of their later achievement in the quantitative realm in elementary school and beyond. Measures of early math skills such as counting, one-to-one correspondence, numeral identification, and solving simple arithmetic problems, are included in several widely-used batteries of early academic achievement, such as the Applied Problems and Quantitative Concepts subtests of the Woodcock-Johnson III Tests of Achievement, the Numerical Operations and Mathematics Reasoning subtests of the Wechsler Individual Achievement Test – II, and the Arithmetic subtest of the Wide Range Achievement Test – 3 (Strauss, Sherman, & Spreen, 2006, pp. 370-400). In the 1998 legislation that reauthorized Head Start, the U.S. Congress mandated that Head Start children know and understand "Numbers and Operations."

► Phonological Awareness

Phonological awareness is another of the skills that are essential stepping-stones on the path to becoming a skilled reader (Whitehurst & Lonigan, 1998; Ehri & Roberts, 2006). It is the ability to segment words into sounds, to understand that spoken words are composed of smaller sound units (root words, syllables and phonemes) (Ehri & Roberts, 2006, p. 114). As Christopher Lonigan (2006, p.78) has written, "Children who are better at detecting and manipulating syllables, rhymes, or phonemes are quicker to learn to read, and this relation is present even after variability in reading skill due to factors such as IQ, receptive vocabulary, memory skills, and social class is partialed out." In the 1998 legislation that reauthorized Head Start, the U.S. Congress mandated that children in the program should acquire phonological awareness.

A difficulty with this domain is that measuring phonological awareness in preschool-aged children is a challenging proposition. Assessment tasks that work well with kindergartners or first graders, such as recognizing rhymes or onset sounds, or telling what word is created when two component words or syllables are blended, or what word is left after a component word or syllable is deleted, show large floor effects and skewed distributions when used with preschoolers, especially those from low-income families. (A floor effect means is said to occur when a majority or large minority of children who take a test score zero.)

► Oral Language Comprehension (Language Screener)

In the interest of not subjecting children to an assessment administered in a language they could not understand, the Office of Head Start decided that the NRS should include a standard procedure for identifying children whose proficiency in English was insufficient for assessment in English. In deciding to go with an English proficiency test, rather than recommendations of parents or teachers, the Office of Head Start followed the procedure recommended by a panel of bilingual experts convened by the National Center for Education Statistics of the U.S. Department of Education. This panel was convened by the design team for the Early Childhood Longitudinal Study of a Kindergarten cohort (ECLS-K) (Montgomery, 1997; Rock & Pollack, 2002, pp. 221-223). The panel noted that states and school districts vary in the criteria they use to identify children's English proficiency and that it was desirable for a large-scale, nationwide study to apply a single standard consistently to all children in the study sample. The panel also noted that the selected measure be relatively short, easy to administer, and easy to score.

The Office of Head Start had an interest in determining the extent to which young children from non-English-speaking families maintained or expanded their knowledge of their home language while they were in Head Start. This was because supporting the home language and culture of English Language-Learners (ELLs) is one of the stated goals of the Head Start program. Bilingual experts advised the Office of Head Start that knowing how proficient ELL children were in their home language would be useful in understanding their progress or lack of progress in English acquisition. For both of these reasons, the Office of Head Start decided that children from families in which Spanish was the first language should receive NRS assessments administered in Spanish, as well as in English. The Spanish battery would also include a language proficiency screener, again in the interest of not subjecting children to an assessment administered in a language they could not understand.

The Office of Head Start determined that it would not be feasible to have parallel NRS assessments and screeners in languages other than Spanish. This was because the number of other language backgrounds represented by children in the Head Start program was quite considerable. It would require a great deal of time, effort, and money to develop parallel assessments in all of these languages. At the same time, many local programs would have great difficulty locating, recruiting and retaining individuals who could administer child assessments in languages other than English or Spanish.

Selection of specific measures. Once the Office of Head Start chose the early achievement domains that would be represented in the NRS assessment, the NRS contractor, Westat, conducted a review of existing measures that might be suitable for tapping the selected domains. Candidate measures had to be technically adequate. That is, there had to be existing evidence that the measures were reliable and valid for use with 4- and 5-year-olds, particularly those from low parent-education, low-income families. The measures also had to be age appropriate in the sense that most if not all children in the target age range showed considerable growth on the measures over an 8- to 12-month period. The measures also had to have minimal floor or ceiling effects among young children. That is, relatively few children in the target age range would receive either the minimum possible score – suggesting that the test was too difficult – or the maximum possible score – suggesting that the test was too easy -- on the selected measure.

In addition to meeting criteria of technical adequacy, candidate measures had to be relatively short, easy to administer, and easy to score. This was so the burden on local programs of assessing all of their kindergarten-eligible 4- and 5-year-old children would not be too great. It was also necessary so that local staff could be trained to administer the assessments accurately in a reasonable period of time. As far as reliability was concerned, measures were sought that had reliability coefficients equal to or greater than .80 . In research or assessment settings where only group means are reported (as in the NRS, where only program-level means were to be reported), any effort to increase internal consistency reliabilities beyond $r_{xx} = .80$ is considered unnecessary (Nunnally & Burnstein, 1994).

Some of the specific measures that were otherwise strong candidates for inclusion in the NRS battery – such as the full version of the Peabody Picture Vocabulary Test, Third Edition (PPVT—III) – had the drawbacks of being relatively lengthy to administer and requiring that the assessor follow one rule about which set of plates to use in beginning the assessment of a given child, then keep track of the child's errors on each set, and follow another rule about where to end the assessment. Such procedures would increase the likelihood of administrative error and also increase the training burden on the Head Start staff to ensure accurate and reliable administration of the measure. Therefore, Westat made use of psychometric methods based on Item Response Theory (IRT) and extensive bodies of data available from national longitudinal studies like FACES and ECLS-K to develop abbreviated and simplified versions of these tests specifically designed for use with 4- and 5-year-old Head Start children.

Each of these abbreviated tests contained a fixed number of items that were selected for optimum discrimination and coverage of the Head Start child population. Each was designed so that all of the items would be administered to every child in the NRS, with no need for the assessor to make decisions about where to begin or end the testing of a given child. Each of the abbreviated tests contained some common bridging items and some items that varied from those contained in the other versions. This made it possible to administer different versions of the test in the fall and spring, while at the same time using IRT methods to equate the difficulty level of the two tests.

National field test. A field test of the HSNRS assessment battery, training, and data collection procedures was conducted in Spring 2003. A national probability sample of 36 Head Start programs, including two migrant and seasonal farm-worker programs and two American Indian programs, was selected, resulting in over 1,430 Kindergarten-eligible English- and Spanish-speaking children being selected for the field test.

The measures included in the field test were the following:

1). An English-Language Proficiency screener, composed of 20 items drawn from the Simon Says and Art Show subtasks of the PreLAS 2000;

2). A Vocabulary task, made up of twenty-four items drawn from the Third Edition of the Peabody Picture Vocabulary Test;

3). A Letter Naming task developed by Westat for the Head Start Quality Research Consortium, consisting of the 26 letters of the English alphabet, presented as upper- and lower-case pairs on three plates, each containing 8 or 9 letters;

4). An Early Math Skills task developed by Westat and composed of items similar to those used in the ECLS-K Mathematics assessment; and,

5). A Phonological Awareness task, consisting of a slightly abbreviated and simplified version of the Elision subtask from the TOPEL Phonological Awareness scale.

A parallel set of measures in Spanish were also included in the field test. These were:

1). An Spanish-Language Proficiency screener, composed of 20 items drawn from the "Tio Simon" and "La Casita" subtasks of the PreLAS Espanol (Duncan & De Avila, 1986);

2). A Spanish Vocabulary task, made up of twenty-four items drawn from the Test de Vocabulario en Imagenes Peabody (TVIP);

3). A Spanish Letter Naming task developed by Westat for the Head Start Quality Research Consortium, consisting of the 30 letters of the Spanish alphabet, presented as upper- and lower-case pairs on four plates, each containing 8 letters;

4). A Spanish Early Math Skills task developed by Westat and composed of Spanish translations of items similar to those used in the ECLS-K Mathematics assessment; and,

5). A Spanish Phonological Awareness task, consisting of a slightly abbreviated and simplified version of the Elision subtask from the Spanish version of the TOPEL Phonological Awareness scale.

The field test results showed that both the English- and Spanish-language versions of the assessment battery had good psychometric properties. The direct assessment measures, in general, showed acceptable internal consistency reliability, especially for program-level reporting. Through the analysis of parallel data collection (where children were assessed by both local Head Start staff and professional data collectors with more experience), inter-assessor reliability was found to be acceptable, except for the phonological awareness Elision task. Another measure that did not perform as well as intended was the Spanish version of the PreLAS Art Show task, La Casita. Assessment means, standard deviations, and reliability coefficients were generally similar whether children were assessed by local Head Start staff or by experienced data collectors. Results were also generally similar whether children were assessed by their own classroom teachers or by other local Head Start staff members. The field test findings played a large part in determining the composition of the assessment battery that was used in the National Implementation of the Head Start National Reporting System.

Moving to national implementation. Guided by the field test results, the Office of Head Start decided that the Language Screener, Vocabulary, Letter Naming, and Early Math measures would be included in the NRS child assessment for national implementation. These components displayed good reliability and performed well with the Head Start children in the field test. The Spanish language version of the assessment battery also performed well. Generally, reliability was good and the battery was appropriate for assessing Spanish speaking children.

However, analysis of the field test data showed lower levels of reliability for the phonological awareness task than for other components of the proposed battery. Further concerns were expressed by Hispanic assessors regarding the Spanish language version of the test. While phonological awareness was recognized as an important building block in early literacy development, the Office of Head Start decided to exclude the Elision task in both English and Spanish versions of the national implementation battery due to these concerns. National Office staff felt that additional time was needed to find or develop a suitable measure for assessment of phonological awareness in preschoolers.

There were also concerns regarding the lower reliability of the Spanish version of the Art Show language-screener task, La Casita. The presentation of the items in La Casita was found to be confusing during the field test, so the presentation of the images was redesigned for the full implementation. The original words used in La Casita were retained, but images appeared alone in picture frames, using the Art Show format. In addition, different regional Spanish names were included in the list of acceptable responses.

With these modifications, the Office of Head Start felt comfortable that the NRS assessment batteries in both English and Spanish were ready to be implemented in local programs throughout the country. At the same time, the OHS committed itself to continuing monitoring of the performance of the batteries as they were implemented in all local programs with all Kindergarten-eligible 4- and 5-year-old children. There were two facets to the continuous monitoring process: One facet was a Quality Assurance study conducted on national samples of Head Start programs and children by an independent contractor. The second facet consisted of continuing analyses and reporting of the psychometric properties of each measure in the assessment battery during each year of national implementation. These results of these analyses were to be published in the technical reports prepared on each year of NRS national results.

Reporting Results To Programs

The NRS produced reports on the average levels of child achievement in the fall and spring of each program year for some 1,800 Head Start programs across the country. Each assessment component was scored separately. There was no attempt to produce a single summary score for the entire assessment. In addition to an overall mean number and percentage of items that children got correct, programs were shown the percentages of children who were at six different skill levels. In the spring, fall-spring growth charts provided information on the gains children had made in particular programs. Local programs could compare their results with national averages and with average results for other programs that had similar demographic characteristics to their own. For example, reference tables and reports were produced for programs from different regions of the country; for programs that had high, moderate, or low proportions of: English Language Learners; children from racial and ethnic minority groups; and children with disabilities. Reference tables were also produced for children in programs in full-day and part-day classes, and those with high, moderate, or low proportions of teachers with Bachelor's Degrees.

To illustrate the kinds of data local programs received from the NRS, Figure 1 shows the average percent of items correct (IRT "true scores") for each NRS English assessment component in the fall of 2005 and spring of 2006. The figure shows national data for all Head Start children who were assessed in both fall and spring. Note that the chart shows substantial growth in each of the assessment areas. In vocabulary and early math skills, children got approximately half the items correct in the fall, on average. By the spring, the children got more than two-thirds of the items correct. In letter naming, the children correctly identified only one-quarter of the letters of the alphabet in the fall, compared to six-tenths of the letters in the spring. Although these were all significant gains, the achievement levels of Head Start children in the spring remained below those of U.S. children from non-low-income family backgrounds. The latter would get all or nearly all of the items in each assessment task correct by the time they were ready to begin kindergarten.

Figures 2, 3, and 4 show fall-spring growth from a different perspective: the percentages of *children* in a program who are below a low skill level, and the percentages who are at or above a high skill level, in the fall and spring. (The figure shows these percentages across all programs in the nation.) The data shown in Figure 2 are only those for children who were English Language Learners. It shows that, based on their performance on the English Language Screener, 38 percent of these children had limited or no understanding of spoken English in the fall, whereas only 12 percent had such limited English proficiency in the spring. At the same time, the proportion of English Language Learners who comprehended directions well rose from 25 percent in the fall to 52 percent in the spring.

Figure 3 shows the vocabulary skill growth for all NRS children who passed the English Language Screener in the fall. It shows that the proportion of Head Start children whose vocabulary skills were one year or more behind age norms went from 54 percent in the fall down to 34 percent in the spring. The proportion of children who were at or above national norms for their age level rose from 12 percent to 31 percent. Again, although there was substantial progress, the vocabulary knowledge of the majority of Head Start children remained below national norms for U.S. children as a whole.

Figure 4 shows the skill level growth chart for the area of letter recognition. It shows that nearly three-quarters of Head Start children could identify 9 or fewer letters of the alphabet in the fall. By the fall, the proportion at that low skill level fell to about one third. This means that, in the spring, two-thirds of Head Start children had achieved the Congressionally-mandated goal of knowing ten letters or more of the alphabet. Indeed, the high skill level data show that the proportion of children who could identify 17 or more letters rose from an 18-percent minority in the fall to a 51-percent majority in the spring.

How programs have made use of NRS findings. NRS reports have been incorporated into the planning and self-evaluation process of many local Head Start programs. NRS data have also been used in local programs' reports to their governing boards and local and state educational authorities. At the national level, NRS data have been incorporated into federal reports on program goals and performance made by the Office of Head Start to the U.S. Office of Management and Budget and the U.S. Congress.

Potential Use of Gain Scores to Compare Program Performance

Program-level data from the first three years of experience with the national implementation of the NRS demonstrate that it is feasible to compare grantee levels of performance by means of gain scores based on child assessments. What the data show about the gain scores are the following:

- There are discernible and statistically reliable differences across Head Start programs in the size of the gains that children make from fall to spring in the skills tested in the NRS assessment (vocabulary, letter naming, early math skills). The differences are sizable enough to be practically important.

- Differences among programs in fall-spring gains are significantly related to their gains in previous years. Between one-seventh and one-fourth of the variation in program rankings by gains can be anticipated from their rankings in the previous one or two years. Letter naming is the skill area in which fall-spring gains are most closely related from year to year.

- Not surprisingly, the average size of the gains that children make in different programs is positively related to where the programs rank in terms of achievement levels in the spring. Between one-tenth and one-fifth of the variation in where programs rank in the spring is associated with the size of their fall-spring gains in the same skill area.

- The size of the gains that children make in a given program is negatively related to the average skill levels at which they entered the program. Programs with students who start out with low skill levels tend to catch up a bit over the course of the year. However, these negative relationships are relatively weak. No more than one-fifth of the variation in program gains can be anticipated from their fall rankings in the same skill area.

- The average sizes of children's gains are related across skill areas. Gains in letter naming and early math skills (average correlation coefficient = .64) and vocabulary and early math skills (average correlation coefficient = .55) are more closely related to one another than gains in vocabulary and letter naming (average correlation coefficient = .38).

- One reason why gains in vocabulary and letter naming are not more strongly correlated is that children in a substantial number of programs make sizable gains in letter naming skills, but only modest gains in vocabulary knowledge.

Despite these encouraging findings regarding the feasibility of making comparisons among programs based on achievement gains, the first three years of NRS data also demonstrate persistent differences in average achievement *levels* across Head Start programs. The NRS data show that:

- There are sizable differences across Head Start programs in the average skill levels with which children enter the program. These differences are related to demographic and socioeconomic differences in the populations that different programs serve. Even though all Head Start programs serve predominantly low income families, there are still variations across programs in average parent education levels, average depth of poverty, and proportions of children from different racial, ethnic, and home language background groups.

- The average skill differences between programs are somewhat reduced by the gains that children make from fall to spring, but they are by no means eliminated. Between half and three-quarters of the variation in where programs rank in the spring can be anticipated from their fall rankings in the same skill area. Vocabulary is the skill area with the strongest relationship between initial and final achievement levels.

- Differences across programs in spring achievement levels are fairly stable from year to year. Between four-tenths and two-thirds of the variation in spring program rankings can be anticipated from their rankings in the previous one or two years. Once again, vocabulary is the skill area in which spring achievement levels are most closely related from year to year.

Whereas many Head Start programs seem to have had some success in bolstering achievement and reducing disparities in letter naming skills, they have had less success in doing the same in the area of vocabulary knowledge. And, as pointed out earlier in this report, vocabulary knowledge at the start of elementary school is predictive of children's academic success in the later years of elementary school and beyond.

Evidence On the Validity of NRS Assessments

Systems that, like the NRS, make use of direct child assessment to monitor the performance of preschool programs are sometimes criticized on the grounds that such assessments are "developmentally inappropriate." Critics also contend that the results of such assessment are unreliable and not predictive of later academic achievement. The following sections of this paper present evidence on the validity of the NRS assessments from recent national implementations of NRS, from the 2003 national field test, and from a study in which NRS data were matched with longitudinal data from the Head Start FACES project. The evidence shows that NRS assessments achieve high rates of child cooperation, satisfy operational criteria for being developmentally appropriate, show satisfactory reliability and predictive validity.

Evidence of high cooperation. In saying that direct assessment is developmentally inappropriate, some critics seem to mean that many young children are not equipped to deal with the demands of a standardized testing situation in which they are required to "show what they know" to an unfamiliar adult (or even to a familiar one) in a limited period of time. According to the critics, many preschool-aged children are intimidated by the testing situation, feeling too shy to speak out even though may know the correct answer to a question or being fearful that they may give the wrong answer and appear foolish. Others have difficulty sustaining attention to testing tasks that go on for many minutes. They may start to move about, not focus on the appropriate pictures and give stereotyped or random responses to questions. Still others do not appreciate the significance or understand the role demands of the assessment situation. They may see it as an opportunity to play or fool around. Or they may simply refuse to cooperate with the assessors' requests.

There is no question that those who would assess 3-, 4-, and 5-year-old children will encounter behavioral issues that are less of a problem with elementary-school-aged children. Young children *are* more variable in their mood and alertness than older children. Extreme shyness in an unfamiliar situation *is* more common among younger children, as are problems with sustaining attention to repetitive tasks over tens of minutes. But assessments of young children can be designed to take account of these developmental realities and include accommodations for variations in children's alertness and task orientation to collect reliable and valid data. As clinical pediatric psychologist W. Douglas Tynan has pointed out, on some days, some 3- and 4-year-old children are so overactive and "off the wall" that it is not even possible for a physician or nurse to measure their height or weight. But that does not mean that the same children cannot be successfully measured on another day. And it certainly does not mean that height and weight are not valid measurements or that trying to ascertain the height and weight of young children is "developmentally inappropriate."

Figure 1.

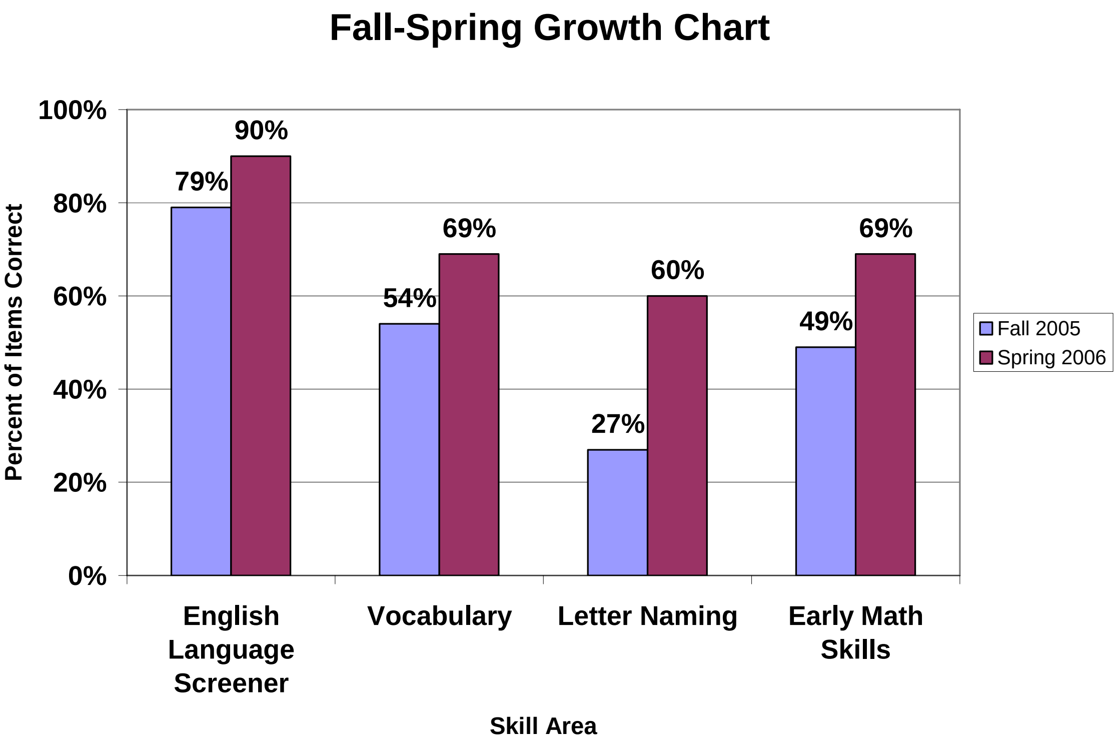

Fall-Spring Growth Chart
100%
80%
60%
40%
20%
0%
Percent of Items Correct
79%
90%
54%
69%
27%
60%
49%
69%
Fall 2005
Spring 2006
English Language Screener
Vocabulary
Letter Naming
Early Math Skills
Skill Area

Figure 2.

Skill Level Growth Chart for English Language Screener (English Language Learners Only)

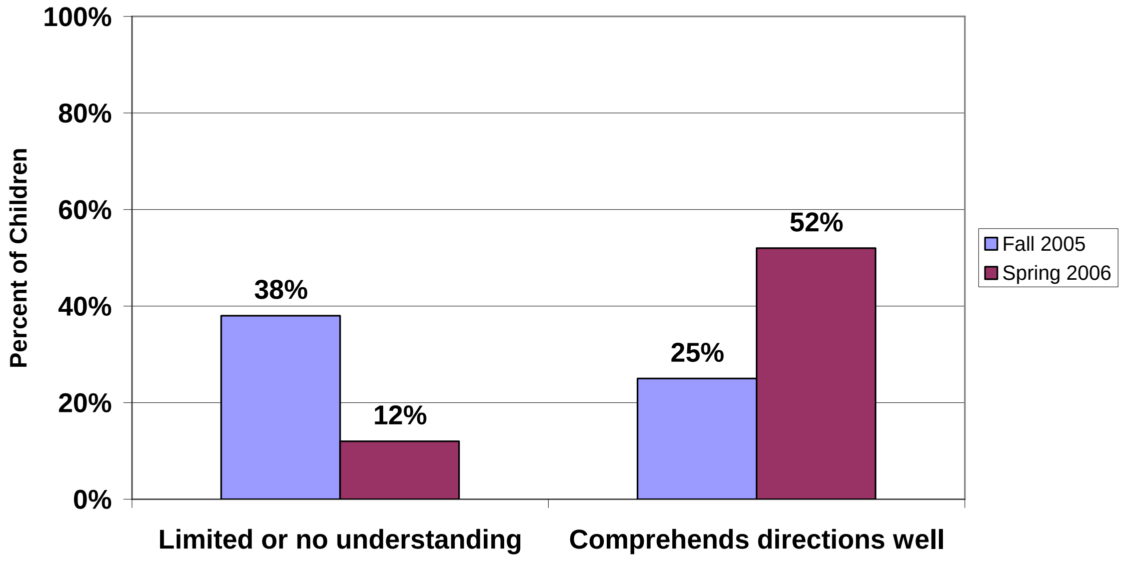


Figure 3.

Skill Level Growth Chart for English Language Vocabulary

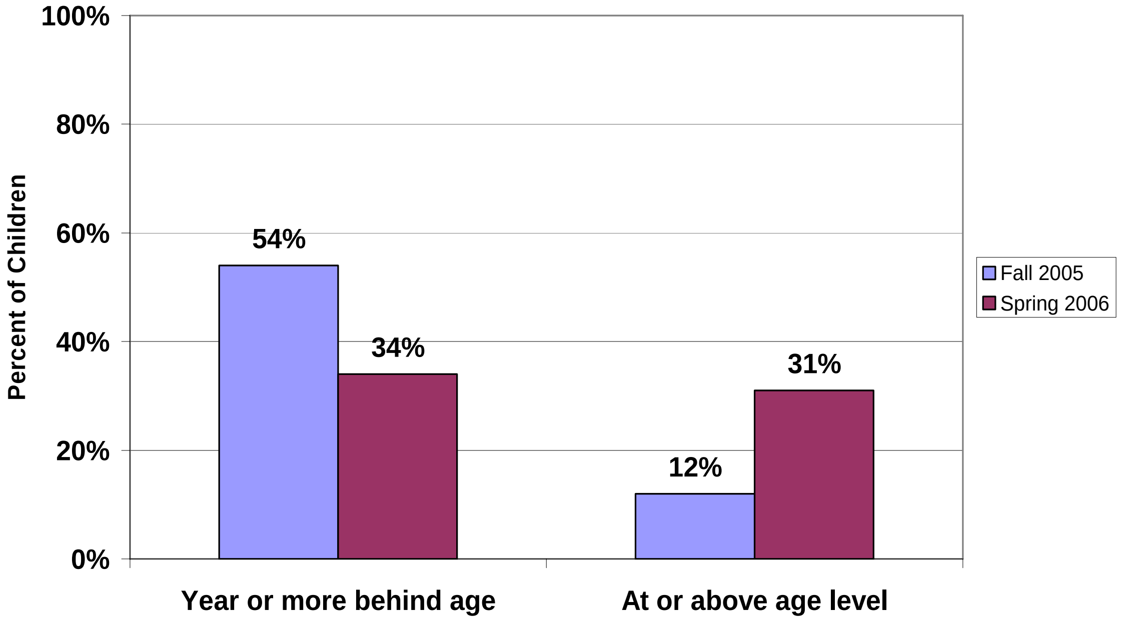
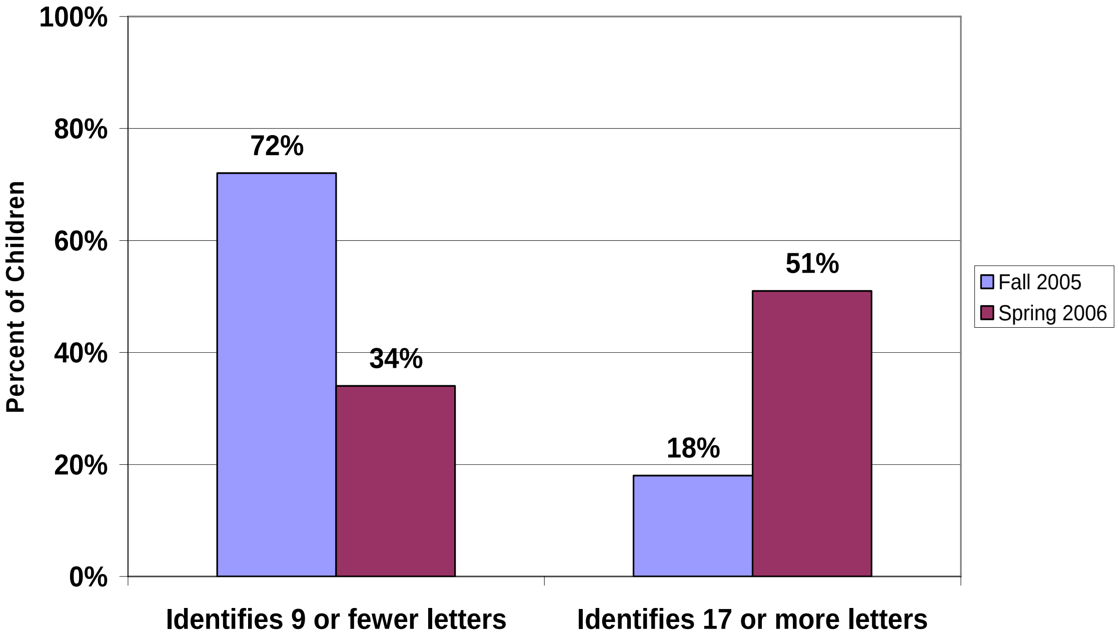


Figure 4.
Skill Level Growth Chart for Letter Recognition (English)



If the problems outlined above were as serious as the critics make out, then it should be the case that attempts to complete a large number of satisfactory young child assessments over a limited time period would meet with substantial failure rates; many of the children would wind up with uncompleted assessments. But the evidence indicates otherwise. For example, in the NRS assessments conducted in fall 2005, during the third year of national implementation, 408,498 children aged 4 and 5 years old were successfully assessed. These children represented 97.3 percent of the children who were eligible to be assessed. Less than one percent – 0.8% – were not assessed because of repeated child non-cooperation. Less than half a percent – 0.4% – were not assessed because of severe disability or because the child's Individualized Education Plan (IEP) prohibited such assessment. And less than two percent – 1.5% – were not assessed because a parent refused permission.

Even among children with diagnosed disabilities and IEPs, 33,866 of 35,796, or 94.6 percent were successfully assessed in the spring 2006 round of the NRS. And, far from being intimidated by the testing situation, the Quality Assurance study of NRS conducted by the independent firm Mathematica found that most local Head Start programs in their national sample reported that children generally responded positively to the assessment experience. Some children appreciated the one-on-one, personal attention from the teacher or other staff member that the assessment afforded.

The difficulty of completing valid assessments with some young children is much more of an issue when child assessment is used to make decisions regarding the academic advancement or placement of individual children than when assessment data are aggregated to form the basis for making judgments about program performance. Some critics of direct child assessment seem not to understand this distinction, or choose to ignore it for the purpose of advancing their argument (e.g., Meisels & Atkins-Burnett, 2004). They react to assessments designed for program evaluation as if they were "kindergarten readiness" tests, even though no such use is intended or made of the assessment data in question.

<u>Evidence that assessment tasks are neither too difficult nor too easy for young children.</u> An indication that a particular assessment task is developmentally inappropriate for preschool-aged children would be if substantial proportions of the children given the task got none of the items right ("floor effect") or all of the items right ("ceiling effect"). The former outcome would indicate that the task was too difficult for most preschoolers. The latter outcome would indicate that the task was too easy for them. Either way, the task in question would not be optimal for showing skill growth during a preschool program year. On the other hand, if most children got a low (but not zero) or mid-range score on the assessment task at the start of the year, and a mid-range to high (but not maximum) score by the end of the year, that would be an indication that the task in question is suitable for showing skill growth in a preschool program. It would also indicate that the task is developmentally appropriate in the sense of tapping a skill and a difficulty range in which most children are showing considerable development during the preschool years.

By these criteria, the assessment tasks that have been used in the Head Start National Reporting System are developmentally appropriate. Both the vocabulary and early math skills tasks showed minimal floor or ceiling effects in either the fall or spring testing occasions. And both tests showed significant growth between fall and spring for children from all ethnic groups and language backgrounds. For example, in the 2005-2006 NRS data, the percent increase in median number of items correct on the vocabulary test was 55 percent for Black children, 46 percent for white children, and 33 percent for Hispanic ELL children. The percent increase in median number of items correct on the early math skills test was 42 percent among Black and white children, and 36 percent among Hispanic ELL children.

The letter naming task showed some floor effect in the fall, with about a quarter of the Head Start children identifying none of the letters of the alphabet. But the same task showed substantial growth in children's letter naming skills between fall and spring. The number of additional letters known by spring by the median Head Start child was 15 among Black children, 12 among white children, and 8 among Hispanic ELL children. Not surprisingly, the English Language Screener task also showed some floor effect in the fall among children who were English Language Learners. More than ten percent of these children had a score of zero on the screener tasks. But this task also showed substantial growth between fall and spring, so that by the spring, less than four percent of the English Language Learners scored zero.

Thus, there is ample evidence that the tasks used in the NRS are developmentally appropriate and can provide indicators of program-related growth in knowledge and skills among preschool-aged children.

<u>Evidence of reliability.</u> There are several test characteristics that one looks at in order to establish the reliability of an assessment task for young children. These include the extent to which a child who does well (or poorly) on one set of items in the test tends to do well (or poorly) on other items in the test. This is called *internal-consistency reliability.* Then there is the extent to which assessments of the same child on two or more occasions tend to produce similar results. (Assuming, of course, that the occasions are not too far apart in time). This is called *test-retest reliability.* Then there is the extent to which assessments of the same child by two different assessors tend to produce similar results. (Again, assuming that the test occasions are not too far apart in time.) This is called *inter-assessor reliability.*

In each round of the national implementation of the NRS, data are analyzed to examine the internal-consistency reliability and other psychometric properties of each of the assessment tasks. These analyses have consistently shown that the abbreviated tasks used in the NRS show reliability comparable to that achieved by trained field assessors in projects like Head Start FACES, the National Head Start Impact Study, or the ECLS-K. During 2005-2006, for example, the median internal-consistency reliability for all NRS English assessment tasks was equal to .84, with a range of .77 to .93. (A coefficient equal to 1.00 would indicate perfect reliability, while a coefficient of 0.00 would indicate complete absence of reliability.) By comparison, the more extended cognitive tasks administered by trained and experienced field assessors in FACES 2000-2001 had a median reliability coefficient of .88 and a range of .77 to .97.

The internal consistency reliabilities for the NRS component tasks were all quite good, though somewhat lower than the highest reliabilities obtained in FACES. All but one of the reliability coefficients was above the .80 level, which is considered appropriate for reporting aggregate scores (Nunnally & Burnstein, 1994). The one score below that level was for the spring English-Language Screener task, in which variation across children was limited because the task was designed to be easy for English-speaking children and most attained high scores on it. (It is difficult to achieve high reliability coefficients when there is limited variation in test scores.)

The NRS 2003 National Field Test described earlier sent trained field assessors to a sample of 36 Head Start programs to conduct parallel assessments of the same children assessed by trained local program staff. All parallel assessments were done within one-to-two weeks of each other, with some of the outside assessments being done before the local assessments and some after. When the parallel assessments were analyzed, the results showed that the inter-assessor reliability coefficient for the NRS tasks had a median value equal to .79, with a range of .78 to .91. While not quite as good as the internal-consistency reliability figures, these reliabilities were quite reasonable, close to the target value of .80.

(The above results do not include the inter-assessor reliability for the TOPEL Elision phonemic awareness task, which was found to have an unacceptably low value (.46). This task was later omitted from the NRS battery used in the national implementations.)

Program-level reliability. The above test reliability statistics are all calculated at the individual child level, which is how test reliabilities are normally analyzed and reported. However, the statistics reported from the NRS are all at the program level, so that the reliability of program mean scores is more germane for program performance monitoring efforts. When child-level assessment scores from NRS were aggregated to produce estimates of mean child achievement levels in fall and spring for each of some 1,800 Head Start local programs around the country, the reliability of the program-level estimates was quite good, generally having values at or above .90.

Reliability of fall-spring growth rates have not been as good, but still respectable, ranging from the low to high .80s. The reliability of estimates based on differences between children's assessment scores at two points in time will generally be lower than the reliability of estimates of mean achievement levels at a single point in time because measurement unreliability at each time point is compounded when difference scores are calculated.

Compared to sample-survey studies like FACES, the NRS is closer in organization and magnitude to what might be required for monitoring the performance of preschool education programs at state or school-district levels. The data indicate that direct child assessment can be done with acceptable reliability in this larger and more challenging context. The NRS results also show that individual child-level assessment scores can be aggregated to produce quite reliable estimates of average achievement levels children attain in different preschool programs.

Evidence of predictive validity. The most potentially damaging criticism of direct assessment of young children, if it were true, is that the results of such assessment are not predictive of children's later achievement. One argument made by critics of direct assessment is that young children's cognitive development does not proceed on a steady trajectory but occurs in fits and starts. A child who appears behind others at one point in time may experience a sudden spurt of cognitive growth that enables him or her to catch up to and even surpass those who were previously more advanced in their development. Thus, assessments at one or two early points in time may have limited predictive value.

A second argument is that early cognitive skills are less important for later school success than non-cognitive aspects of child development such as curiosity, attention span, task persistence, and self-regulation. According to this argument, if assessments of young children are to be truly predictive, they must include these non-cognitive aspects of children's development (Shonkoff & Phillips, 2000; Raver, 2004; Raver & Zigler, 1997, 2004).

Earlier longitudinal studies based on relatively small samples of the young child population (often samples of convenience) seemed to lend some support to these arguments (Horn & Packard, 1985; Stevenson & Newman, 1986; Pianta & McCoy, 1997; LaParo & Pianta, 2000) . However, more recent studies that have followed large, representative samples of young children from preschool into elementary school and beyond have shown that cognitive measures taken at or just before school entry are indeed predictive of later achievement (Zill et al., 2003, Chapter VII; Storch & Whitehurst, 2002; Duncan et al., 2005; O'Donnell & Zill, 2006).

For example, various studies using the ECLS-K longitudinal data have shown that direct assessment results at the start of kindergarten correlate with children's tested performance in reading and math at the end of the first, third, and fifth grades. Assessment results also correlated with the likelihood of being retained in grade through fifth grade. Moreover, the predictive importance of early math skills and general knowledge (the latter correlating highly with vocabulary scores) increases as the child goes from the earlier to the later grades of elementary school.

A limited evaluation of the predictive validity of NRS assessment scores was carried out by matching data for a sample of 671 children who were assessed in both the NRS, while they were in Head Start, and in FACES 2003, at the end of their kindergarten years. There was an 86 percent response rate in the FACES Kindergarten follow-up. However, for purposes of confidentiality, children in the national NRS data were only identified by the Head Start program they were in and their birth dates. This limited the extent to which data could be matched: the effort resulted in a good but not great matching rate of 71 percent (671 out of a possible 942 children from FACES who were assessed in kindergarten). On the other hand, analysis of the matched and unmatched cases did not show any substantial demographic or test score differences between them.

Analyses of data on these matched cases were carried out to examine how well the NRS assessment results correlated with the children's tested performance at the end of kindergarten. The predictive power of the NRS scores was also examined by seeing how well the scores with ratings of the child's academic performance by the kindergarten teacher, and with the teacher's report that the child would be recommended for promotion to first grade or for retention in kindergarten or placement in a "transitional" or "developmental" class.

Multiple regression analyses were performed, predicting from a weighted linear combination of the NRS assessment scores to the kindergarten outcomes. The analyses showed that the NRS scores did indeed have predictive validity. The multiple correlation coefficients were equal to .56 for a basic reading skills composite test score at the end of kindergarten, .60 for a math reasoning composite test score, and .66 for a general knowledge composite test score. The same kind of analysis showed that the NRS scores showed a good relationship with teacher ratings of children's academic performance in kindergarten. For example, the multiple correlation coefficient with ratings of language and literacy skills was .54.

A slightly weaker but still substantial relationship was found when multiple logistic regression analysis was used to examine the relationship between NRS scores and the likelihood that the child would be promoted to first grade or recommended for retention in kindergarten. The logistic regression equivalent of a multiple correlation coefficient was equal to .45, and the percent agreement between the predicted and actual promotion decision was equal to 81 percent. Thus, the NRS assessments have been shown to have predictive validity both with respect to later direct assessments of children and to teacher evaluations of children's academic performance in elementary school.

Summary and Conclusions

This essay has described the Head Start National Reporting System as a possible model for systems aimed at assessing and monitoring the performance of preschool programs. The NRS used direct assessment of a limited set of early language, literacy, and math skills in order to compare different local programs on the average levels of achievement children attain and the average gains children make from fall to spring. The NRS used the "train the trainer" method to train and certify large numbers of local program staff to carry out one-on-one child assessments of some 400,000 children in a reasonably reliable and accurate manner. A Computer-Based Reporting System (CBRS) was used to register 4- and 5-year-old children who are eligible to be assessed and to enter demographic data on the characteristics of programs, centers, classes, teachers, assessors, and children. Children's responses were recorded by local assessors on scannable answer sheets and transmitted to central scanning and analysis facilities. IRT scoring methods were used to equate different versions of the assessment tasks across years and from fall to spring in the same year.

Reports on the average levels of child achievement were produced in fall and spring for some 1,800 Head Start programs across the country. In the spring, fall-spring growth charts provided information on the gains children made in particular programs. Local programs could compare their results with national averages and with average results for other programs that had similar demographic characteristics to their own. NRS reports were incorporated into the planning and self-evaluation process of many local Head Start programs. NRS data were also used in local programs' reports to their governing boards and local and state educational authorities. At the national level, NRS data have been incorporated into federal reports on program goals and performance made by the Office of Head Start to the U.S. Office of Management and Budget and the U.S. Congress.

There are good practical reasons for preferring direct assessments of children's skills to the most widely used alternative method for assessing children's developmental progress, namely, the use of teacher ratings based on observation of the child during ongoing learning and play activities. Measures based on direct assessment are easier to compare across groups of children who attend different preschool programs or attend no program at all. Methods for determining the reliability and validity of direct assessment measures are better developed and easier to implement than procedures for determining the reliability and validity of observation-based measures. Procedures for training staff members to carry out direct child assessments and certifying that they can do so with acceptable accuracy are easier and less time-consuming than are similar procedures for observation-based ratings.

The principal criticisms that have been leveled against direct assessment of young children's knowledge and skills in general, and the NRS assessments in particular, are that such assessments are "developmentally inappropriate," unreliable, and not predictive of later achievement. The evidence marshaled above shows these charges to be invalid. Certainly, there are challenges involved in assessing children in the 3-to-5 year-old age range. But such challenges can be and have been overcome. This is especially true when the focus is on aggregating individual assessments to get a sense of the average progress that children make in one program as opposed to another.

The evidence shows that most children of ages 4-, and 5 can be successfully assessed when the assessment is suitably designed, kept to reasonable length for young attention spans, and administered by properly trained assessors. Though not all children with diagnosed disabilities can be assessed, most can. It is perfectly feasible to develop sets of assessment items that are developmentally appropriate in the sense of being neither too hard nor too easy for preschool children to respond to correctly. Appropriate items are those which typically show substantial growth during the preschool years in the percentage of children who answer them correctly.

When assessments of preschool-aged children are properly constructed and administered, assessment results show adequate internal consistency reliability and reasonable inter-assessor reliability. These reliabilities generally improve as results are aggregated from the individual child to the class, center and program levels. Assessments of vocabulary and general knowledge, early reading and math skills show strong predictive validity in forecasting both children's early and later achievement in elementary school. These results show that the National Reporting System merits consideration as a possible model for those seeking to implement a system aimed at assessing and monitoring the performance of preschool programs in a given state, city, county, or school system.

REFERENCES

Administration for Children and Families. (2001). *Head Start FACES: Longitudinal findings on program performance. Third progress report.* By N. Zill, G. Resnick, K. Kim, R. Hubbell McKey, C. Clark, S. Pai-Samant, D. Connell, M. Vaden-Kiernan, R. O'Brien, & M. A. D'Elio. Washington, DC: U.S. Department of Health and Human Services.

Administration on Children, Youth, and Families. (2003). The Head Start Leaders Guide to Positive Child Outcomes: Strategies to Support Positive Child Outcomes. Washington, DC: U.S. Department of Health and Human Services.

Administration for Children and Families. (2003). *Head Start FACES 2000: A whole-child perspective on program performance.* By N. Zill, G. Resnick, K. Kim, K. O'Donnell, A. Sorongon, R. Hubbell McKey, C. Clark, S. Pai-Samant, R. O'Brien, & M. A. D'Elio. Washington, DC: U.S. Department of Health and Human Services.

Administration for Children and Families. (2005). *Head Start Impact Study: First year findings.* Washington, DC: U.S. Department of Health and Human Services.

Biemiller, Andrew. (2006). Vocabulary development and instruction: A prerequisite for school learning. Chapter 3, pp. 41-51 in: David K. Dickinson and Susan Neuman, Eds., *Handbook of early literacy research: Vol. 2.* New York: The Guilford Press.

Bowey, J.A. (1995). Socioeconomic status differences in preschool phonological sensitivity and first-grade reading achievement. *Journal of Educational Psychology, 87,* 476-487.

Brigance, A.H. (1991). *Brigance Diagnostic Inventory of Early Development (Birth to Seven Years) - Revised.* North Billerica, MA: Curriculum Associates, Inc.

Bureau of Labor Statistics (2004). NLSY79 Child & Young Adult Data Users Guide: A Guide to the National Longitudinal Survey of Youth 1979.

Campbell, J.M., Bell, S.K., Keith, L.K. (2001). Concurrent validity of the Peabody Picture Vocabulary Test – Third Edition as an intelligence and achievement screener for low SES African American children. *Assessment, 8*(1), 85-94.

Chall, J.S. (1983). *Stages of reading development.* New York: McGraw-Hill.

Chew, A.L. (1981). *The Lollipop Test: A Diagnostic Screening Test of School Readiness-Revised.* Lake Worth, FL: Humanics Publishing Group.

Cunningham, A.E., & Stanovich, K.E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33,* 934-945.

Duncan, G.J., Dowsett, C.J., Brooks-Gunn, J., Claessens, A., Duckworth, K., Engel, M., Feinstein, L., Huston, A.C., Japel, C., Klebanov, P., Magnuson, K., Pagani, L. and Sexton, H. (2005). School Readiness and Later Achievement. Paper presented at biennial meetings of the Society for Research on Child Development, April 10, 2005.

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

Duncan, S., & De Avila, E. (1986). *Pre-LAS*. Monterey, CA: CTB/McGraw-Hill.

Dunn, L.M. & Dunn, L.M. (1997). *Examiner's manual for the Peabody Picture Vocabulary Test Third Edition.* Circle Pines, Minn.: American Guidance Service.

Ehri, L.C. & Roberts, T. (2006). The roots of learning to read and write: Acquisition of letters and phonemic awareness. Chapter 9, pp. 113-131 in: David K. Dickinson and Susan Neuman, Eds., *Handbook of early literacy research: Vol. 2.* New York: The Guilford Press.

Fischel, J.A., Storch, S.A., Spira, E.G., & Stolz, B.M. (2003). Enhancing emergent literacy skills in Head Start: First year curriculum evaluation results. Presentation at the Biennial Meeting of the Society for Research in Child Development, Tampa, FL.

Ginsburg, H.P., & Baroody, A.J. (1983). *The test of early mathematics ability.* Austin, TX: Pro-Ed.

Glover, M.E., Preminger, J.L., Sanford, A.R. (1988). *Early LAP: The Early Learning Accomplishment Profile for Young Children: Birth to 36 months.* Chapel Hill, NC: Chapel Hill Training-Outreach Project.

Harlaar, N., Dale, P.S., & Plomin, R. (2007). From learning to read to reading to learn: Substantial and stable genetic influence. *Child Development, 78(1),* 116-131.

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore: Brookes.

Hart, B. & Risley, T. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator, 27(1),* 4-9.

Horn, W.F., & Packard, T. (1985). Early identification of learning problems: A meta-analysis. Journal of Educational Psychology.

LaParo, K.M. & Pianta, R.C. (2000). Predicting children's competence in the early school years. A meta-analytic review. Review of Educational Research, 70, 1373-1400.

Lonigan, C.J., Wagner, R.K., & Torgesen, J.K. (2007). *Test of Preschool Early Literacy Examiner's Manual.* Austin, TX: Pro-Ed.

Lonigan, C. J. (2006). Conceptualizing phonological processing skills in prereaders. Chapter 6, pp. 77-89 in: David K. Dickinson and Susan Neuman, Eds., *Handbook of early literacy research: Vol. 2.* New York: The Guilford Press.

Magnuson, K.A., Ruhm, C., & Waldfogel, J. (2004). *Does prekindergarten improve school preparation and performance? National Bureau of Economic Research Report.*

Mardell-Czudnowski, C. & Goldenberg, D. S. (1998). *Developmental Indicators for the Assessment of Learning – Third Edition (DIAL–3): Manual.* Circle Pines, MN: American Guidance Service.

Meisels, S.J. & Atkins-Burnett, S. (2004). The Head Start National Reporting System: A critique. *Young Children.*

Montgomery, D. (1997). *Identification of an English proficiency measure for the Early Childhood Longitudinal Study.* Palo Alto, CA: American Institutes for Research.

Nehring, A. D., Nehring E. F., Burni, Jr., J. R., & Randolph, P. L (1992). *Learning Accomplishment Profile – Diagnostic (LAP-D) Standardized Assessment: Technical Report – 1992 Revision and Standardization.* Lewisville, NC: Kaplan Press.

Newcomer, P.L., & Hamill, D.D. (1997). *Test of Language Development (3rd Ed.).* Austin, TX: PRO-ED, Inc.

Nunnally, J.D. & Burnstein, I.H. (1994). *Psychometric theory (3rd Ed.).* New York: McGraw-Hill.

O'Donnell, K., & Zill, N. (2005). Predicting early and later reading achievement in children from low-income families from skills measured at kindergarten entrance. Poster presented at the Annual Meetings of the Population Association of America, Los Angeles, CA, April 2006.

Pianta, R.C., & McCoy, S.J. (1997). The first day of school: The predictive validity of early school screening. Journal of Applied Developmental Psychology, 18, 1-22.

Raver, C.C. (2004). Placing emotional self-regulation in sociocultural and socioeconomic contexts. Child Development, 75, 346-353.

Raver, C.C., Smith-Donald, R., Hayes, T., & Jones, S.M. (2005, April). Self-regulation across differing risk and sociocultural contexts: Preliminary findings from the Chicago School Readiness Project. Paper presented at the biennial meeting of the Society for Research in Child Development, Atlanta, GA.

Raver, C.C., & Zigler, E. (1997). Social competence: An untapped dimension in evaluating Head Start's success. Early Childhood Research Quarterly.

Rock, D.A., & Pollack, J.M. (2002). Early childhood longitudinal study – kindergarten class of 1998-99 (ECLS—K): Psychometric report for kindergarten through first grade. Washington, DC: U.S. Department of Education: National Center for Education Statistics, Report No. NCES 2002-05.

Sanford, A. R. & Zelman, J. G. (1995). *The Learning Accomplishment Profile – Revised Edition.* Chapel Hill, NC: Chapel Hill Training-Outreach Project.

Shonkoff, J., & Phillips, D. (Eds.) (2000). From Neurons to Neighborhoods: The Science of Early Childhood Development. Washington, DC: National Academy Press.

Starkey, P., Klein, A., & Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Research Quarterly, 19(1),* 99-120.

Stevenson, H.W. & Newman, R.S. (1986). Long-term prediction of achievement and attitudes in mathematics and reading. *Child Development, 57*, 646-659.

Storch, S.A., & Whitehurst, G.J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38,* 934-947.

Strauss, E., Sherman, E.M.S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary. Third edition.* New York: Oxford University Press.

U.S. Department of Health and Human Services, Administration for Children and Families. (2005, January). Report to Congress: Assessment of Children in Head Start Programs. Washington, DC: Author.

Washington, J.A., & Craig, H.K. (1999). Performances of at-risk, African-American preschoolers on the Peabody Picture Vocabulary Test—III. *Language, Speech, and Hearing Services in the Schools, 30,* 75-82.

Wasik, B.A., & Bond, M.A. (2001). Beyond the pages of a book: Interactive book reading and language development in preschool classrooms. *Journal of Educational Psychology, 93(2),* 243-250.

Weizman, Z.O., & Snow, C.E. (2001). Lexical input as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology, 17,* 265-279.

Whitehurst, G.J., & Lonigan, C.J. (1998). Child development and emergent literacy. *Child Development, 68,* 848-872.

Williams, K.T., & Wang, J.J. (1997). *Technical references to the Peabody Picture Vocabulary Test – Third Edition (PPVT—III).* Circle Pines, Minn.: American Guidance Service.

Woodcock, R. W. & Johnson, M. B (1989). Woodcock-Johnson Revised Tests of Achievement. Itasca, IL: Riverside Publishing.

Zill, N., & O'Donnell, K. (2006). The frequency of grade repetition from kindergarten through fifth grade among students from low-income families: Analysis of longitudinal data from ECLS-K. Rockville, MD: Westat.

Zill, N., & Resnick, G. (2006). Emergent literacy of low-income children in Head Start: Relationships with child and family characteristics, program factors and classroom quality. Chapter 25, pp. 347-375 in: David K. Dickinson and Susan Neuman, Eds., *Handbook of early literacy research: Vol. 2.* New York: The Guilford Press.

Zill, N., & West, J. (2001). Findings from the *Condition of Education 2000*: Entering kindergarten. Washington, DC: U.S. Department of Education, National Center for Education Statistics. Publication No. NCES 2001035.