## Measuring Readiness of Head Start Children

**Chair:** David J. Armor
**Discussant:** Donald A. Rock
**Presenters:** Nicholas Zill, C. Cybele Raver, Christine Li-Grining, Kyle Snow

- **The Use of Direct Assessment of Children's Cognitive Skills in the Evaluation of Preschool Programs**
  Nicholas Zill

- **Disentatangling Components of School Readiness**
  Kyle Snow

- **Direct Assessments of Children's Emotional and Behavioral Skills**
  C. Cybele Raver, Christine Li-Grining, Radiah Smith-Donald
  (text unavailable)

**Zill:** A growing number of states are providing publicly-funded preschool programs, along with increased calls for programs to be held accountable for achieving measurable results. The discussion is how the contribution of these programs to children's development should be evaluated. Zill's position is that direct assessment of children's cognitive skills should be an integral part of program accountability efforts and that growth in cognitive skills should not be the sole focus of the evaluation process, but it should be at least a major component of a comprehensive approach to program evaluation.

It is necessary to focus on early cognitive skills because they are important for children's later achievement, and stakeholders and policy makers care about growth in cognitive skills. Another reason for this focus is that valid procedures for assessing cognitive skills are readily available, while procedures for assessing social-emotional development are less well developed, making them more challenging to implement in evaluation efforts. The same drawbacks are true for assessments of children's self-regulation and executive functioning.

The alternative approaches to measuring cognitive development are teacher ratings based on observation of children during ongoing learning and play activities, or direct assessment of children's cognitive skills and early academic knowledge. Problems with observation-based ratings are:

1. Establishing reliability and comparability of ratings by different teachers is challenging. Independent raters must 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 activities. In practice, reliability and comparability of ratings may not be established, which undercuts the value of the evaluation.

2. Training and certifying teachers to do ratings in a reliable and standardized manner is also challenging. The trainee and instructor must observe children on a number of different occasions over a relatively extended period of time, while the children are engaged in a

variety of ongoing classroom activities. Then they must compare ratings; also burdensome and difficult to arrange. In practice, teachers may be allowed to do ratings without adequate training and certification, which undercuts the value of the ratings.

3. Teacher ratings tend to have psychometric limitations and defects. One such problem is the Halo effect where the teacher forms a global judgment of the child's academic capabilities. Such ratings are derived from this overall impression, rather than observations of specific skills. A teacher may use a sliding scale when rating children from different ethnic groups, language groups, or SES backgrounds. Also, better-trained teachers may rate more strictly. This can lead to a paradoxical finding that children with better prepared teachers show less progress.

There are also criticisms of direct assessment: behavioral issues prevent many children from being assessed; direct assessments of young children are not developmentally appropriate; results are not reliable or valid. There are several sources for rebutting these criticisms, such as the Head Start Family and Child Experiences Survey (FACES), the Head Start National Reporting System (NRS), and the Early Childhood Longitudinal Study of a Kindergarten Cohort (ECLS-K).

Possible behavioral issues in assessing young children are that the child is too shy to respond, even though he or she knows the correct answer, or that the child is fearful of giving the wrong answer or appearing foolish. With an overly active child or one with a short attention span, the assessor would have difficulty keeping the child focused on the assessment task. Another problem is that the child may become discouraged as questions become harder, and he or she stops trying.

While young children are more challenging to assess than are older children, assessments can be designed to minimize behavioral issues, and assessors can be taught strategies for dealing with behavioral issues. Successful assessment of every child is not required for program accountability purposes, but evidence shows that a vast majority of children in major national projects are successfully assessed. These high cooperation rates were achieved in the fall 2004 NRS, FACES, and ECLS-K. Many children not only cooperate with assessment, they enjoy the personal one-on-one attention they receive.

Operational criteria for developmental appropriateness insure that the tasks are not too difficult for children in the relevant age range (floor effect) or that the tasks are not too easy for the children in the relevant age range (ceiling effect), and that the typical child shows growth in task performance during the program year. Developmental appropriateness for children in the 4- to 5-year-old range is evidenced by the fact that most tasks used in FACES and all tasks used in NRS show substantial growth in task performance during the program year. The same growth was shown by ethnic minority children and English-language learners, as well as by nonminority children.

Percent increase in median raw score from fall to spring (2004-2005) show the following: NRS Vocabulary equals 31 percent for Black children, 27 percent for White children, and 30 percent for Hispanic English-learners. The percent increase in NRS early math skills for the same period are 42 percent for Black, 31 percent for Whites, and 36 percent for Hispanic ELL children.

For children in the 4- to 5-year-old range, most tasks were neither too difficult nor too easy. There were minimal floor and ceiling effects, except for the letter-naming task, but this task shows a large growth during the program year. The NRS letter naming task found the following: for Black children the median raw score was four letters, with 20 percent identifying no letters in the fall; for White children it was two letters, and 36 percent; and for Hispanic ELL the numbers were two letters and 18 percent identifying no letters. Black children could identify 15 letters by spring, with 12 for Whites, and 8 for Hispanic ELL children.

Reliability is established by: (a) internal consistency, where children who do well on one component of a task tend to do well on other components; (b) test-retest reliability, where assessments of the same child on two occasions produce similar results; and (c) inter-assessor reliability, where assessments of the same child by two different assessors product similar results. Evidence from FACES 2000-2001, NRS 2004-2005, and the 2003 NRS field test show good reliability in all three areas.

To establish predictive validity, assessment results must correlate with the child's tested performance at the end of kindergarten and in later grades. In addition, assessment results must correlate with teacher ratings of the child's academic performance in kindergarten and later grades, and assessment results must correlate with school administrative decisions (e.g., promotion to first grade). Preliminary evidence shows predictive validity among a sample of children assessed in both NRS and FACES during 2003-2004. NRS assessment results correlate with the child's tested performance at the end of kindergarten, which predict basic reading skills, math reasoning composite, and general knowledge composite. NRS assessment results correlate with kindergarten teacher ratings of the child's academic performance at the end of kindergarten NRS assessment results also correlate with a school administrative decision to have a child repeat kindergarten or attend a transitional class.

Longitudinal follow up of children in ECLS-K study show that direct assessment results at the start of kindergarten correlate with children's tested performance at the end of first, third, and fifth grades. Assessment results correlate with the likelihood of being retained in a grade through fifth grade. The predictive importance of early math skills and general knowledge increases in later grades.

In conclusion, children's cognitive skills at the end of preschool and entrance to elementary school are predictive of both early and later achievement. Direct assessment of children's early skills can be done with developmental appropriateness, reliability, and validity and direct assessment has practical advantages over observation-based ratings. Some preliteracy skills (such as letter knowledge and decoding skills) are predictive of basic reading achievement; whereas others (vocabulary and general knowledge) are predictive of reading comprehension. Math skills predict both. Direct assessment of children's cognitive skills should be an integral part of preschool program accountability efforts.

**Snow:** The National Education Goals Panel defined the domains of school readiness as: (a) physical well-being and motor development, (b) social and emotional development, (c)

approaches toward learning, (d) communication and language usage, and (e) cognition and general knowledge. The School Readiness Act of 2005 includes in its own definition:

1. language knowledge and skills, including oral language and listening comprehension;
2. prereading knowledge and skills that prepare children for early literacy in schools, including phonological awareness, print awareness, and alphabetic knowledge; and
3. premathematics knowledge and skills, including aspects of classification, seriation, number, spatial relations, and time;
4. social and emotional development related to early learning, school success, and sustained academic gains;
5. and in case of limited-English proficient children, progress toward acquisition of the English language while making meaningful progress in attaining the knowledge, skills, and development described in 1 through 4 above.

Regardless of the source of definitions, the two defining features that remain constant are that school readiness is comprised of multiple skill sets and capacities which vary within the population, and that these skills and capacities are both independent and interrelated. To date, the research literature has focused predominantly on the first of these themes—identifying key competencies and demonstrating variation within the population, describing these at school entry, and developmentally to include precursors and later outcomes. The second theme, examining interrelations of key school readiness components has been less frequently explored.

These interrelationships are important to understand because they help to identify subgroups of children with different profiles in order to target intervention or instruction. Interrelationships are also important in the development of assessments that minimize cross-domain confounding, as well as for modifying interventions.

Three studies: (a) Colvig-Amir, Liu, and Mobilio (2005); (b) Hair, Halle, Terry-Humen, and Calkins (2003); and (c) Konold and Pianta (2005) have examined interrelations of school readiness domains using cluster analysis. These studies applied cluster analysis to large samples of children at the time of kindergarten entry. They all used measures of multiple domains, and all examined sociodemographic differences between identified clusters. These studies showed predictive relations between cluster membership and later outcomes.

Colvig-Amir, et al. surveyed the kindergarten teachers in Santa Clara County, CA using an instrument designed to provide data on the relative proficiency levels of the children on 20 items drawn from the framework provided by five domains identified by the National Education Goals Panel ($n$=943). The children were put into four readiness groups: the All-Stars, Needs-Prep, Social-Stars, and Focused-on-the-Facts.

The All-Stars group (48%) was reported as near or at proficiency in all five NEGP domains; Needs-Prep (11%) was reported at not quite in progress on attaining any of the five NEGP domains; the Social-Stars (15%) were rated high on emotional and social development, physical well-being and motor development, but are not as prepared in communication and language usage, cognition, and general knowledge domains. The Focused-on-the-Facts children (26%) are near proficient in cognition and general knowledge, but below proficiency in other areas.

The clusters in this study differed in family income, delaying kindergarten enrollment until the child is at least 5 years old, preschool experience, proficiency in English, English spoken as the primary language in the home, parents reading to their children, and parental education. For example, 62% of All-Stars came from families with incomes over $82,000, as opposed to 8% of Needs-Prep and Social-Star families. More than three in four All-Star families speak English in the home, as compared to less than 40% of Needs-Prep and Social-Star families. A large number of All-Star and Focused-on-the-Facts families read to their children at high levels, which is true for only two-thirds of the Needs-Prep and Social-Star families.

Hair, et al. examined child competency using data from the ECLS-K cohort in four domains: health, social and emotional functioning, emergent literacy, and general cognitive functioning. They collapsed the scores in each domain, using proficiency scores derived from raw data provided by direct assessment of children, and parent and teacher report; then constructing clusters based upon risk relative to children's expected levels of proficiency, The cluster characteristics in this sample are in the following areas: health risk (low birth weight, disabilities, teen parents); social and emotional risk (low birth weight, not in two-parent home, low parent education, teen parent); language or cognitive risk (low parent education, teen mother, non-English home); low risk (two-parent home, higher parental education, mother married at child's birth, English spoken at home).

Results found that in the health domain the children were nearly 2.5 standard deviations below the population mean in aspects of physical health, including rating of general health, body mass index, and fine and gross motor skills. In the domain of social and emotional risk the children scored nearly 2 standard deviations below the mean in areas such as self-control, social interaction, and behavior problems. For the language and cognitive risk domain the children scored greater than a 0.5 standard deviation below the mean on skills such as emergent writing and reading and problem solving, classifying, and sorting. In the low risk domain the children scored greater than the mean in all areas, including nearly 1 standard deviation above the mean in the language and cognition measures. The findings predicted that in first grade there would be differences in reading, math, ratings of "works to best of ability," and ratings of self-control.

The Konold and Pianta study applied cluster analysis to a group of cognitive and social developmental measures among 54-month-old children in the NICHD Study of Early Child Care and Youth Development. Their profile found attention problems (10 %), low cognitive ability (7%), low to average social and cognitive skills (20%), social and externalizing problems (17%), high social competence (24%), and high cognitive ability and mild externalizing (22%). Clusters in this study differed in mother's age, mother's education level, partner's education level, income-to-needs ratio. Concurrent differences were found between clusters in PPVT scores, letter-word identification, and Woodcock-Johnson applied problems. In 1$^{st}$ grade, cluster differences were present in the same three measures.

These three studies provide some convergence in findings despite different data sources and measures. There is a "ready for school" group that meets most or all expectations. These clusters carry both expected predictive relations and socio-demographic predictors based on the literature. Also, there are blended strength (or risk) clusters that have differences in their

readiness in social versus cognitive skills. The level of maternal education, and family income (variably measured) were common correlates of cluster membership.

In the future, research is needed to determine if the clusters are stable over time and if there are identifiable transitions between clusters. Future research should also determine if cluster membership is subject to intervention effects and whether clusters can be replicated using different measures and samples.