Early Childhood Education and School Readiness:
Conceptual Models, Constructs, and Measures

Washington D.C.
June 17-18, 2002

WORKSHOP SUMMARY

Workshop Sponsors:
U.S. Department of Health and Human Services
National Institute of Child Health and Human Development
Administration on Children, Youth and Families
Office of the Assistant Secretary for Planning and Evaluation

TABLE OF CONTENTS

Early Childhood Education and School Readiness Workshop:
Conceptual Models, Constructs, and Measures

I. INTRODUCTION

On June 17-18, 2002, a multidisciplinary group of experts was convened to advise the National Institute of Child Health and Human Development (NICHD), the Administration on Children, Youth and Families (ACYF), and the Office of the Assistant Secretary for Planning and Evaluation (ASPE) within the U.S. Department of Health and Human Services (DHHS) on the measurement and assessment of learning and development in early childhood, and on priorities for measures development. Participants included experts in early language, literacy, and mathematics; cognitive, social, and emotional development; regulation of attention, behavior, and emotion; early school achievement and transition; special education; reading disabilities; developmental cognitive neuroscience; research design and quantitative methods in intervention, longitudinal and observational research; and early childhood policy, practice, and professional development (see Appendix A for list of participants).

This meeting was the first in a series to inform a set of research and programmatic initiatives funded by the DHHS and the U.S. Department of Education. The intent of the initiatives is to stimulate research and apply scientific knowledge to ensure that all children, from birth through age five, develop the early knowledge, skills, and competencies needed to benefit from high-quality instruction in kindergarten and the early grades. Specifically, an Interagency Early Childhood Research Initiative funded by the DHHS and Department of Education has been announced to encourage research that answers the overarching question: Which early childhood programs or combinations of program components and interactions with adults and peers are effective or ineffective in promoting early learning and development, for which children, and under which conditions? The results will be used to inform early childhood programs and practices in pre-kindergarten, home-based and center-based child care, family childcare, and Head Start.

In addition, the President's early childhood initiative, *Good Start, Grow Smart,* charges the Head Start Bureau under the ACYF with designing a national reporting system to monitor the progress of every child in Head Start in the legislatively mandated areas of language, literacy, and early numeracy. The data will be used for program improvement and accountability. Piloting of assessments was scheduled for fall 2002 with full implementation set for fall 2003. This meeting also informed that effort.

Workshop presenters and participants were asked to consider current research on what children should learn and develop from birth through age five to prepare for kindergarten and the early grades. For both research and programmatic purposes, they were asked the questions: What constructs should be measured? What are the strengths and weaknesses of available instruments? What approaches should be taken in developing a set of instruments with adequate coverage? What are the priorities for instrument development and research to support the development of new measurement and assessment tools? Individual working groups focused first on constructs and measures within specific domains of development (i.e., language and literacy; cognition and mathematics; regulation of attention, behavior, emotion; social-emotional competency). Next, multidisciplinary groups recommended strategies for developing measurement packages that assess children's progress across all domains, and outlined priorities for future instrument development and research. Part One of this document summarizes these presentations and discussions.

Part Two describes sessions that focused specifically on Head Start's national reporting system. Discussions laid groundwork for a July 9, 2002, meeting that was sponsored by the ACYF in collaboration with the NICHD to finalize assessment and design options. Though the focus was on Head Start, the ideas generated were valuable for informing any systematic attempt to design, implement, evaluate, and report on the effectiveness of large-scale, comprehensive, early childhood education programs. In addition, discussions revealed gaps in available and widely used measurement and assessment tools, and in basic knowledge of learning and development that limit the potential of such a system for program evaluation and improvement. These gaps present clear challenges to research communities across disciplines. Thus, Part Two contains additional priorities that emerged in discussions for developing new measures and generating the knowledge needed to design systems that would be most useful for evaluating and promoting children's progress.

This summary report is being disseminated to share the content of workshop discussions. The report is not intended to provide a comprehensive review of the relevant scientific knowledge base, an exhaustive inventory of measures, or complete descriptions of the measures that were discussed. Citations are provided for additional information, where possible. The report is also not intended to be an endorsement of particular approaches to measurement. The constructs, measures and priorities for measurement development included reflect the specific interests and expertise of the workshop participants. The event was not a consensus conference, however; and thus the content should not be interpreted as representing the views of each participant. Presentations and discussions were recorded, synthesized into a draft report, and participants were given the opportunity to review the draft and correct factual errors. The co-sponsors of this event thank the participants for the time and effort they devoted to discussions and to reviewing and editing these proceedings.

II. PART ONE: CURRENT STATE OF KNOWLEDGE AND PRIORITIES FOR INSTRUMENT DEVELOPMENT

A. OVERVIEW

To provide context and background, presentations by experts in early childhood research, practice, and policy described a set of current approaches to measuring early learning and development in studies of early childhood interventions, large-scale nationally representative surveys of school readiness, impact research on Head Start and Early Head Start, descriptive child care research, classroom-based observational assessments of children's progress, and a 50-state study of early childhood program standards and assessment systems. Presenters explained measurement strategies, discussed strengths and weaknesses of existing tools, highlighted new tools under development, and suggested areas where new measures are urgently needed. Each presentation was followed by a period of open discussion.

Several general recommendations emerged from presentations and group discussions, motivated by strengths and weaknesses of approaches to measurement that experts observed in the field:

- Ground instruments in child development theory and data
- Develop measures that have practical relevance
- Use measures appropriate for the population (e.g., norms and psychometrics should be appropriate to the language, culture, age-span of the group studied; clinical measures are not ideal for studying typical development)
- Include direct child assessments with parent and teacher report
- Use and develop measures with sound psychometric properties (while avoiding ad-hoc, a-theoretical construction)
- Establishing and following guidelines for training and administration
- Controlling for type I and type II errors and repeated testing effects
- Developing efficient and integrated systems of assessment that provide robust measures across all areas of learning and development and that are based on the most recent and rigorous scientific findings

In group discussions, participants recommended that measurement strategies for research be designed to achieve breadth and depth, including both a broad sampling of items across content domains as well as a set of in-depth measures that comprehensively assess constructs within each domain of interest and that are tailored to the intervention. Participants suggested constructs that should be covered in specific areas of language, literacy, early mathematics, social and emotional development, and regulation of attention, behavior, and emotion, and directions for research and instrument development in each content domain.

Researchers discussed the strengths and weaknesses of existing instruments. Of particular interest were measures that could be used or developed as core instruments and that would allow the combining of data across studies to answer larger sets of research questions, comparing of results, and conduct of meta-analyses. As the instruments were discussed, researchers categorized them into one of three tiers according to whether or not they are well-standardized, promising experimental measures, or appropriate for use in large-scale research relating to early childhood interventions or programs. A few measures that enjoy popularity were not recommended for continued use, especially in studies of early childhood intervention.

B. CONTEXT AND BACKGROUND

Presenters with expertise in research, practice, and policy outlined goals and criteria for selecting measures, strengths and weaknesses of existing instruments, challenges of assessing and reporting on children's progress, and suggestions for developing new measures and assessment systems. Themes that emerged across presentations and working groups are summarized together in the next section, *General Suggestions for Measurement, Instrument Developmen,t and Research*.

Susan Landry, University of Texas-Houston, Health Sciences Center

Dr. Landry presented approaches to measurement used in research to implement and evaluate the effect of a model professional development program for Head Start teachers. Designed by researchers and educational training staff from the Center for Improving the Readiness of Children for Learning and Education (CIRCLE), the model is based on research supported with over 15 years of federal, state, and private funding on factors most important for supporting young children's cognitive and social development. The program promotes social and emotional growth while focusing on three key program components critical to later reading and academic success: 1) language development, 2) early literacy (i.e., phonological awareness, letter knowledge, written expression, book and print awareness, motivation to read), and 3) early mathematics (e.g., number and operations).

JoAnn Robinson, University of Colorado, Health Sciences Center

Dr. Robinson described theoretically motivated approaches to selecting constructs and measures for three recent intervention studies: **Home Visitation 2000; Memphis New Mothers Study; and Early Head Start-Denver**. The interventions span the period from birth through age seven. The studies were designed to test the hypothesis that inadequate prenatal care and dysfunctional styles of parent-child interaction impair neurobiological growth and negatively affect emotion and behavior regulation, as well as cognitive and executive functions, resulting in poor academic performance and antisocial behavior. Thus, interventions that ensure adequate prenatal care and more functional parent-child interactions are predicted to provide children with early experiences needed to benefit later from strategies intended to promote early literacy, language, and cognitive development.

John Love, Mathematica Policy Research, Inc.

Dr. Love described the **Early Head Start Research and Evaluation** project in which families were randomly assigned to Early Head Start or to a control group. The project includes outcome measures for infants and toddlers from 3,000 low-income families living in 17 diverse communities in the United States (http://www.acf.dhhs.gov/programs/core/ongoing_research/ehs/ehs_intro.html).
Dr. Love emphasized that though measures in all areas need development, better specification of constructs and instruments are especially needed to assess approaches to learning (task persistence, curiosity, initiative, planfulness); social and emotional development; language development (to assess more than receptive vocabulary); and motor development.

Dr. Love proposed that a system of measurement be developed to meet the challenge of comprehensively measuring the effectiveness of early childhood programs that would be feasible for use in large-scale research and that would assess children's progress across time. The system would include a collection of standard measures that minimize time, cost, and training, contain optimally balanced content and methods (direct standard assessments and report measures), and a standard administration order. A function of the system would be to assess sets of outcomes that correspond to goals that should be the focus of early

childhood programs regardless of setting (preschool and Head Start classrooms, center- or home-based child care programs, and family child care). Data on environmental factors that support or impede progress would be essential for interpreting the results and making judgments about how best to allocate resources to improve programs. Knowledge of multiple normative developmental trajectories, less typical trajectories not predictive of risk, and trajectories indicating risk for poor outcomes, would be incorporated into the system and be used to evaluate progress. Though parts of such a system could be designed using the existing scientific evidence-base, more complex, multi-level, longitudinal data on diverse populations of children are needed to meet the challenge of developing a system with all of these characteristics.

Nicholas Zill and Ronna Cook, Westat Inc.

Dr. Zill presented an overview of the **Head Start Family and Child Experiences Survey (FACES)**, a longitudinal study of program performance that compares child outcomes to national norms (http://www.acf.dhhs.gov/programs/core/ongoing_research/faces/faces_intro.html). Data collection in the second cohort begins in fall 2002 with a sample of 2,800 children entering 43 Head Start programs. A multi-method approach to measurement includes direct assessment, observation, parent and teacher reports, and behavior ratings. Direct assessments are obtained at the beginning and end of the Head Start year and at the end of kindergarten. Those administering assessments receive one week of training with quality control follow-up. For the assessment battery, subscales or items were selected from standard normed measures and unnormed assessments created to cover social, emotional and language development, early literacy, cognitive development and general knowledge, motor development and physical well-being, and approaches to learning. (For a list of measures, see http://www.acf.hhs.gov/programs/core/ongoing_research/faces/faces2000_instruments/face2000_intro.html).

Ms. Cook presented proposed measures for the **National Head Start Impact Study**, which begins in fall 2002 (http://www.acf.dhhs.gov/programs/core/ongoing_research/hs/impact_intro.html). The study involves 75 grantees, and 5,000 to 6,000 children who will be randomly assigned to Head Start or to a control group. The current proposal is to build upon the FACES battery, retaining or adapting measures that previously showed improved performance. Goals for improving assessments are to strengthen the oral language assessment of the battery, select more measures that assess growth over time, and ensure trained field interviewers can administer the measures with acceptable reliability. Criteria for selecting measures included: absence of major floor or ceiling effects with Head Start populations, measures that predict school achievement, appropriate length to maintain interest and performance levels, and availability of parallel tests in Spanish and English, at least for a subset of the battery.

Jerry West, National Center for Education Statistics, U.S. Department of Education

Dr. West provided an overview of the **Early Childhood Longitudinal Study-Birth Cohort (ECLS-B)** that follows children born in 2001 from birth through the end of kindergarten, and the **Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K)** that follows children from kindergarten entry through the end of the fifth grade year (http://www.nces.ed.gov/ecls/). The goal of these longitudinal studies is to report how well public schools, early childhood programs, and other environments in this country prepare children for school and otherwise affect their lives. A specific goal for the birth cohort is to assess factors that are believed to be precursors of later development and scholastic success, and that link conceptually to the battery used to assess children's progress in the kindergarten and first grade. The lack of national standards for the knowledge and skills children are expected to develop in kindergarten and elementary school complicated selecting constructs and instruments to track children's progress from

birth through age five with respect to school readiness. Research on child development and school achievement led to the selection of constructs in the areas of physical development (e.g., height, weight, middle upper-arm circumference) motor skill, language, literacy, mathematics, general knowledge, prosocial and problem behavior, emotion-regulation, self-control, and approaches to learning. Most standard measures were inappropriate because of inadequate coverage or developmental range resulting in the adaptation and development of new measures.

Dr. West emphasized the urgent need to develop measures to assess progress across the period of early childhood that, when combined, have sufficient breadth and depth of coverage across multiple domains and yet remain feasible in terms of time, cost, training, and implementation on a large-scale. Studies that seek nationally representative samples require instruments that are appropriate for assessing all children. Though instruments in most areas need development, direct assessments of social development and social skills are among the most limited.

Margaret Burchinal, University of North Carolina-Chapel Hill

Dr. Burchinal presented measures used in the **NICHD Study of Early Child Care and Youth Development** (http://www.nichd.nih.gov/od/secc/index.htm). Since 1991, a diverse sample of 1,300 families from 10 sites throughout the country has been followed in the study to determine the relation between children's early experiences and developmental outcomes. Dr. Burchinal presented data analyzed for this meeting on associations between cognitive, language, and social-emotional measures collected during early childhood and later measures of reading achievement and social skill in first grade. For the period of birth through three, measures included the Bayley Mental Development Index; the MacArthur Child Development Index (CDI); the Achenbach Child Behavior Checklist (CBCL); the Adaptive Social Behavior Inventory (ASBI); the Bracken School Readiness; and the Reynell Auditory Comprehension and Expressive Language Scales. In first grade, children received the Woodcock Johnson Academic Composite and selected subtests (i.e., letter identification, applied problems, incomplete words and memory for sentences); and the Preschool Auditory Comprehension and Expressive Language Scales.

Sharon Lynn Kagan, Teachers College, Columbia University

Dr. Kagan, in collaboration with Dr. Catherine Scott-Little, is undertaking an analysis of standards for early childhood programs and assessment systems used in the 50 states. The impetus for setting standards and assessments differs across states and may originate with governors, state legislators or members of the early childhood community. Preliminary data show that 28 states have standards of learning and development. Though many states fund pre-kindergarten programs, no state has a state-wide strategy for assessing all children and few states have standards that align with the assessment system. Approaches taken by South Carolina, North Carolina, and Michigan were described to illustrate differences in purpose, design, standards, and instrumentation. A nomenclature problem must be remedied, which means that targets of measurement (constructs) must be defined and relations among them described. This knowledge must be incorporated into early childhood systems in the form of frameworks and benchmarks that align with the content of assessments used to guide classroom practices and evaluate program effectiveness.

General parameters must be set for designing assessments and reporting systems. Ideally, standards for programs serving four-year-olds would be aligned with standards for kindergarten, which in turn would be aligned with first-grade. Technical requirements for measures differ according to the purpose of

assessment. A recommendation was that all instruments and systems should be useful for program improvement. The purposes of the Head Start national reporting system are to improve quality of instruction in Head Start classrooms, and to monitor trends for program evaluation and accountability. Any system designed to assess children for reporting at the federal level must consider how it would be aligned with systems at the state level, which differ in design and purpose within states.

Next steps for Head Start and other early childhood systems throughout the nation include: a) ensuring the appropriate selection of assessments depending on the purpose; b) generating better information on cost, implementation, training and other supports needed to collect reliable and valid information and to use it effectively to improve programs; and c) piloting existing instruments and developing new instruments because most research measures do not meet these practical needs.

Sarah Brainerd, Stamford Connecticut and Sharon Lynn Kagan, Teachers College, Columbia University

Ms. Brainerd directs two Head Start programs, one center-based and one mixed center- and home-based, that are participating in a researcher-practitioner partnership with Columbia University. Goals for the 2001-2002 academic year were to utilize the Program Review Instrument for Systems Monitoring (PRISM), a national assessment tool for reporting on Head Start programs (see Part Two of this document), and meet legislatively mandated outcomes for Head Start. Challenges have included identifying measures for assessing children's progress that teachers can use to develop the curriculum and to support children's development, identifying measures appropriate for the Spanish-speaking and bilingual children and staff, and supporting parents in their abilities to promote children's progress. The program is currently piloting a portfolio-based, work sampling system in two classrooms. Trained staff will help individual classrooms use the assessments and review data on program strengths and areas for improvement.

Gayle Cunningham, Head Start, Birmingham, Alabama and Martha Abbott-Shim, Georgia State University

Ms. Cunningham oversees Head Start and Early Head Start programs with 21 locations in public schools, and a budget that allows only for hiring paraprofessionals. Ms. Cunningham described challenges in reviewing Head Start Performance Standards and selecting tools appropriate for assessing children's progress, with a priority on measuring the 13 legislatively mandated outcomes in literacy, language, and numeracy. The Chicago Early Learning Tool was among a variety of tools that were examined and rejected for inadequate coverage. The Language Early Learning Assessment was created in partnership with Dr. Abbott-Shim to measure developmental progress and to generate profiles that will be used for individualizing the curriculum and guiding group instruction. Researchers emphasized this measure is appropriate only for staff development and planning, and may be aggregated only at the classroom level. Reliability and validity have not been evaluated.

C. GENERAL SUGGESTIONS FOR MEASUREMENT, INSTRUMENT DEVELOPMENT, AND RESEARCH

Presenters, working groups, and larger discussions produced the following general suggestions for selecting, using, and developing measures of children's learning and development for large-scale research on comprehensive early childhood programs.

Achieve Breadth and Depth

Two sets of instruments are needed:

1) A broad sampling of items across content domains that give best indicators of children's status, and

2) A set of in-depth measures that comprehensively assess constructs within a domain and that are tailored to the intervention.

The following constructs were recommended for coverage in the content areas that were a primary focus of this workshop. Participants also emphasized the importance of assessing gross, fine, and visual motor development and physical health.

- **Language.** Many studies include only measures of vocabulary, typically assessed with the Peabody Picture Vocabulary Test, which is limited to receptive vocabulary. Though receptive vocabulary is important to assess, it is only one part of language development, and scores should not be interpreted as giving information about children's general language abilities. Assessment of children in the preschool period should include receptive and expressive language as well as language context. The specific areas to be measured include: narrative, orthographic knowledge, phonology (of each language the child speaks or is learning), syntax (including relational knowledge and morphology), discourse and pragmatics, verbal repetition, and metacognitive foundations. Language context includes naturally occurring family and classroom language and intentional language scaffolding. For comprehensive assessment, perinatal hearing screening, infant speech discrimination, and motor control related to speech production are recommended.

- **Early Literacy.** Known precursors and predictors of later reading achievement should be assessed. These include: phonological processing, immediate and longer-term phonological memory, phonological sensitivity, phonological awareness, print awareness, decoding, receptive vocabulary, general knowledge, letter knowledge, and inventive spelling.

- **Mathematics.** Instruments limited to numeracy do not adequately measure mathematics content knowledge and skills that should be promoted during the period of early childhood. The following areas are important to assess from preschool through second grade: number sense, number operations (e.g., counting, arithmetic that includes addition and subtraction for ages three to five, algorithms and strategies); geometric reasoning and spatial cognition (e.g., shapes, locations, transformations and symmetry); measurement, patterns, and data; processes that include problem-posing and solving, reasoning, and communicating, and understanding of cross-cutting concepts (e.g., part-whole composition/decomposition). Different aspects of mathematical thinking and behavior are important to assess, including concepts, procedures, and metacognitive processes, such as those that motivate checking one's work or choosing appropriate methods of solution (see the Conference on Early Math Standards Web site, http://www.gse.buffalo.edu/org/conference/ for a more complete description of constructs and known developmental sequences that should be the targets of instruction and measurement). Also see the Web site of the National Association of the Education of Young Children for their joint position statement, along with the National Council of Teachers of Mathematics, concerning early childhood mathematics instruction, http://www.naeyc.org/.

- **Social-emotional competency.** Key constructs include: emotion regulation (e.g., regulation of positive emotion, negative emotion, and delay of gratification); emotion expressiveness, social engagement; pro-social skills and cooperation (which have cognitive, social, and empathic or emotional subcomponents); absence of behavioral problems, such as externalizing (aggressive) and internalizing behavior (depressed or withdrawn), hyperactivity, and anti-social behavior; social cognition (includes understanding of emotions of self and others) and social pragmatics.

- **Regulation of attention, behavior, and emotion.** Key constructs recommended for measurement include: sustained attention, inhibitory control, working memory, activity level, intrinsic motivation/mastery, planning, problem-solving, goal setting, task persistence, curiosity, engagement, automaticity of cognitive processing and execution of responses, flexibility of attention, behavioral flexibility (associating behaviors with particular settings and not others), and engagement with materials. The development of executive functioning underlies many of these (inhibitory control, planning, goal setting, working memory, persistence, flexibility, and perhaps others). There was disagreement about the importance of measuring compliance (e.g., execution of requested behaviors, negotiations of requests, and responsiveness to adults with appropriate vocalizations and gestures). Overly compliant or dependent behavior can be associated with poor cognitive and social outcomes; yet the ability to adapt behaviors flexibly in response to requests made by adults or peers predicts academic achievement and social functioning in group-based settings, such as classrooms and other early childhood settings.

Ground Instruments in Theory and Developmental Data

Measures selected should be based on recent theory and constructs selected should be precursors of the short- and long-term outcomes that are the goals of the early childhood program. The ad-hoc and a-theoretical construction of measures (e.g., extracting 4 or 5 items from a larger instrument without first validating the adapted measures) results in untrustworthy data.

Instruments should track growth over time, and give markers of progress in each domain of interest (e.g., cognitive, social and emotional, mathematics, language and literacy development, physical development and health), without floor or ceiling effects. Ideally, measures should provide continuous assessment of progress from birth through age six years. To do this with existing measures, instruments would most likely have to be aligned across two periods: infancy (birth to 18 or 24 months), and the toddler/preschool period through kindergarten (18-24 months to 5-6 years). Though instruments should generally assess the same constructs within a domain across ages, the elements within a domain that are important to measure often change across ages, especially in the areas of language, pre-literacy, mathematics, and cognition. Thus, all elements may not need to be measured longitudinally.

Develop Measures that Have Practical Relevance

Some research measures have high predictive validity and other sound psychometric properties but lack face validity, mostly because their applied relevance is not clear. Practitioners in attendance indicated that measures with face validity are better received at the local program level because teachers can "see" the importance of an assessment. How sensitive decontextualized and normative global measures should be to intervention is not clear. For example, though the Bayley Scales of Infant Development is widely regarded as a gold standard measurement, some researchers recommended the measure should not be used as the primary tool for evaluating the effects of early childhood programs, especially if the intention is to use the data for program improvement. Similarly, the implications of research findings obtained using decontextualized measures, such as the Leiter Assessment of Sustained Attention, for intervention

and instruction are not always apparent. The approach of using standardized measures to account for variance in the effect size of interventions needs to be complemented with a criterion-referenced approach that can show where children are in a developmental sequence of knowledge and skills, and document how much children have progressed across time or in response to intervention. Few such measures exist and to develop them, normative longitudinal data must be collected to fill gaps in knowledge concerning precursors, developmental milestones and trajectories, and the combination of individual, environmental, social, and neurobiological factors that influence their development.

To adjust instructional strategies and improve the effectiveness of programs, teachers need assessments for determining which children have mastered particular skills. Continual, dynamic assessments need to be developed and the education and professional development required for administering reliable and valid assessments needs to be specified. The content of dynamic assessments should align with curriculum goals and benchmarks and with measures used in research to evaluate program effectiveness.

Use Measures Appropriate for the Population

Norms and psychometric data for existing measures and for new measures must be obtained for diverse samples that represent the demographics of U.S. children and families. Large-scale studies provide an opportunity to obtain this information. Problems with existing instruments, such as floor and ceiling effects, need to be eliminated to make them sensitive measures for children across a larger developmental range.

Some commonly used and highly regarded measures developed for use in clinical settings may not be appropriate for assessing the development of broader populations of children. When measures developed for clinical purposes are used with broader populations, behavioral growth often occurs in these populations that goes undetected. For example, some participants reported that the Achenbach CBCL, which was developed for clinical use, meets gold standards for measurement but often does not appear sensitive for measuring intervention effects for broader populations. Similarly, measures used to screen for risk typically should not be used to measure the effect of early childhood programs on typical development.

The child's native language and dialect must be considered when selecting, using, or developing new measures. Bilingual assessments cannot be performed exclusively in one language or the other, or separately in both languages. A child may know a portion of the concepts in each language, suggesting that a composite or merged score from measures conducted in each language should be used; however, identical assessments generally cannot be obtained in both languages because simple translations do not make equivalent measures. The choice of measures and strategies for developing composite scores depend on the goal of assessment and the type of information sought. Researchers should consider implications of allowing code switching or language switching within a test.

Cultural sensitivity must be considered when selecting constructs and instruments. Differences in cultural norms and values (e.g., Asian and U.S. Caucasian values of independence, sustained attention, and so on) have implications for setting goals for early childhood education programs and selecting measures to assess outcomes. Though a consensus was not reached on this issue, the following suggestions emerged: Most behaviors (e.g., self-regulatory behaviors) are important for human functioning in a variety of cultures, but the contexts for displaying these behaviors, and the conditions that elicit them (or not) may differ. Research is needed to determine how cultural differences in environments and interactions affect individual differences in capabilities and developmental trajectories. Ultimately, decisions about measurement probably depend in part on the purpose of the intervention. That is, if the goal of intervention is to prepare diverse populations of children for formal schooling, which assumes a common

set of elements, then the definition and measurement of desired outcomes may not differ across children or programs. A desirable approach would be to operationally define a set of core expected outcomes, assess whether cultural differences moderate effectiveness of the intervention, and if so, determine how and why.

Include Direct Child Assessments with Parent and Teacher Report

Studies should include multiple measures from multiple perspectives that consist of direct standard assessments of children as well as teacher and parent report. Behavior is often context-specific, making it difficult to determine what a child knows or can do with a brief assessment conducted at a specific point in time.

The quality of standard measurement increases with age because children become less inhibited and more responsive. Thus, parent and teacher measures may in some cases result in better information and more complete data (e.g., when assessing their own children). Yet, parent and teacher reports may reflect characteristics or biases of the respondent, and teacher reports may be biased according to child characteristics that include but are not limited to culture, ethnicity, race, and gender. Moreover, discriminations among children tend to improve with teacher education, and teachers with more years of experience tend to give children higher ratings. In addition, precautions should be taken to ensure teacher-rating tools do not lose sensitivity, which can occur when teachers rate every child in a classroom.

Report measures should be validated using standard direct assessments; however, this suggestion is complicated by the lack of standardized instruments for directly assessing children, especially during infancy and in the areas of social, emotional, and behavioral development. There is a particular need to develop standard direct assessments that adequately cover essential constructs in all domains across the period of early childhood.

Use and Develop Measures with Sound Psychometric Properties

Participants agreed that all measures should meet high standards for reliability and validity. Validation studies should be conducted on adapted measures in order to establish their psychometric properties even if the measures were developed using rigorous and defensible procedures. The consensus was that predictive validity, construct validity, content validity, and concurrent validity are all important and should not be compromised. For example, some participants expressed concern that the development of resource-efficient measures that are feasible for large-scale research can lead to documenting the predictive validity of adapted measures that may lack content and construct validity based on sound theory and/or previous research.

All measures should be developed to meet similar accepted psychometric standards, but any inequalities in psychometric soundness should be considered when interpreting results. Unless all measures selected meet similar high standards, results showing that some measures have greater predictive power than others may reveal more about uneven measurement development than about child development and the factors that influence it. Finally, the criteria used to establish the predictive validity of measures used in research on early childhood programs should be both theoretically and practically meaningful and stated explicitly.

Establish and Follow Guidelines for Training and Administration

When selecting, using, or developing new instruments, it is critical to determine whether characteristics of the examiner affect the results (whether a stranger, gender, ethnicity, match with child's demographics). Educational background, type, and intensity of training or certification required for obtaining reliable and valid data should be determined. Standards for education and training may vary widely depending on the measure, child population (developmentally delayed or learning disabled) and purpose of measurement. For example, measures of articulation and phonological representation are important for measuring language functioning, but the educational background or training needed to distinguish between articulation and phonological difficulties should be specified.

Control for Type I and Type II Errors and Repeated Testing Effects

When analyzing large data sets containing many measures, stringent controls for Type I and Type II errors must be applied. Most repeated measures designs suffer from learning effects, and therefore, research designs should protect against obtaining higher scores in the absence of real change. Another option following Item Response Theory (IRT) is to generate assessments using items selected randomly from a large universal pool, which would also avoid inflated scores that could result from "teaching to the test." This procedure would provide a low-cost, science-based method of developing assessments that teachers can use to assess progress, and that researchers can use to evaluate the effectiveness of intervention and instruction.

Develop Efficient and Integrative Systems of Assessment

Most research measures are labor-intensive and require specialized expertise or labor-intensive training. More efficient measures are needed to comprehensively assess learning and development in the context of large-scale data collection. Burdens of time, cost, and training must be reduced while ensuring the collection of reliable and valid data across a range of settings (e.g., preschools, center-based and home-based programs, family day care). When designing comprehensive assessments, compromises must be made about which constructs to include. Attempts at comprehensive coverage can lead to the a-theoretical and ad-hoc construction of instruments that potentially compromise psychometric soundness, and thus, the trustworthiness of the data.

One option for developing more efficient, theoretically based measures, is to develop standard procedures for integrated direct assessments. A small set of tasks would be used to collect data on multiple components of a single domain, such as language, or across multiple areas. Promising methods or paradigms for developing integrated assessments include: narrative elicitation (story telling and re-telling), observations of problem-solving (e.g., social, early mathematics), teaching paradigms (e.g., Nursing Child Assessment Satellite Training [NCAST]: Teaching Task Scales), and methods used to assess joint attention during infancy. Underlying processes shared across domains would be assessed as well as component knowledge and skills that are unique to each domain. To this end, reliable and valid neurobiological techniques may be developed for use with preschoolers during passive behavioral tasks to identify processes such as metacognitive or relational processing, working memory, and verbal memory and other processes hypothesized to underlie learning and development across domains. Integrated assessments would minimize redundant variation and add value to domain-specific measures by assessing processes and skills underlying development across multiple domains believed to be associated with

currently unexplained variance in outcome measures. Thus, standard measures would be more efficient, comprehensive, and yield more complete and interpretable results.

D. SUGGESTIONS FOR INSTRUMENT DEVELOPMENT AND RESEARCH IN CONTENT DOMAINS

Language and Early Literacy

In addition to comprehensively measuring individual components of language (e.g., syntax, receptive vocabulary), it is essential to determine how children use language to express themselves and to solve problems. Thus, integrated assessment procedures should be developed to assess language in contexts that allow assessment of social pragmatics and cognitive processes. Metacognition and relational processing, working memory, and verbal learning are underlying processes that should be assessed across domains. Studies are needed to determine the relation between language and literacy development, and to identify shared metacognitive or metalinguistic precursors.

The degree of independence between pre-academic domains during the preschool years must be determined to better design prevention strategies that can prevent later difficulties in reading and mathematics. It is known that problems with reading and math tend to co-occur in the early grades, though the association is not perfect. Floor and ceiling effects and other confounds must still be ruled out, but the co-occurrence points to a shared underlying process that potentially affects performance in both domains. Participants noted that much of the variance in reading and mathematics performance remains unexplained, and research is needed to determine if a small constellation of shared processes may be accounting for problems that occur in both areas. If these processes were identified, then early childhood prevention strategies could target these as part of a more comprehensive approach to preventing later reading and mathematics difficulties. To pursue this line of prevention research, better measures are required to assess the development of reading, mathematics, and hypothesized shared underlying processes from preschool into the elementary grades.

Phonemic awareness is a precursor of reading ability, but causes of the failure to develop phonemic awareness are not known. To detect early precursors of phonemic awareness, which would allow for focused intervention, better measures of phonological sensitivity must be developed that do not require sophisticated behavioral, social, or linguistic responses and that enable, for example, assessment of sensitivity to speech contrasts during infancy. Research is needed to further develop the use of evoked response potentials (ERPs) into standard assessments of early speech discrimination skills, speech sensitivity, and language sensitivity, which have been shown to predict reading ability.

In the literacy field, the construct of phonological short-term memory has been measured using a variety of methods that include non-word (pseudo-word) repetition tasks, digit span, sentence repetition, and word span (e.g., Wagner & Torgeson, 1987). The construct is included in standard literacy measures for older children (e.g., Comprehensive Test of Phonological Processing [CTOPP]; Wagner, Torgesen, & Rashotte, 1999). Adaptations of the CTOPP are under development for three- and four-year-olds (Lonigan, Wagner, Torgesen & Rashotte, in preparation). However other experimental measures of phonological short-term memory for use with preschoolers, especially those which have been shown to predict academic success and to be sensitive to schooling effects, also should be further developed for use in research and applied settings.

Promising basic research paradigms may be useful for developing standardized infant assessments, including conditioned head-turn, head-turn preference, non-nutritive sucking and anticipatory heart-rate

response procedures. Statistical learning in infants is a promising area of research with possible implications for early childhood education and intervention, but more work is needed that bridges the gap between basic and applied research methods to establish the utility of these methods. Interventions that promote children's capacity to engage in joint attention have led to improvements in language development, specifically in vocabulary and reference skills. The methods used in these interventions should be further developed into standard assessments that would be useful for identifying children for intervention, designing interventions, and evaluating intervention effectiveness. Neurobiological techniques that could be used in conjunction with behavioral tasks include Magnetic Resonance Imaging (fMRI), structural MRI, magnetoencephalography (MEG), and Diffusion Tensor Imaging (DTI).

Measures of language development collected during story telling and retelling (e.g., Frog Stories; see Berman, Ruth A., & Slobin, D. I., in collaboration with Aksu-Koc, A. A. et al., 1994) predict later academic achievement. Research is needed to develop this widely used method into standard assessments. The tasks are easy to administer, but have labor-intensive procedures for collecting, coding, scoring, and analyzing data. Informal protocols exist for coding and scoring data, but a more standard approach for administration, coding, and analysis is needed. Assessments could be obtained in multiple languages, and a single task yields multiple quantitative measures that include but are not limited to phonology, expressive language, morphology, syntax, language complexity, and narrative structure.

Language measures tend to be inappropriate for addressing complex issues of dialect and second language that arise, for example, when Mexican children mix Mexican dialect with Puerto Rican or African American English dialects. Most measures do not have other-language versions or scoring options that take dialect into account. Adaptations of existing measures must be made for speakers of languages other than English, and new measures developed. Language outcome measures are needed for use with bilingual children and for African American children who speak a dialect of English that has syntactic, vocabulary and phonological differences that can affect communication and literacy development. When assessing speakers of languages other than English, goals for language assessment must be explicit and the selection of measures made according to whether the goal is to assess vocabulary development or vocabulary in a particular language, and the basis upon which decisions to test in one language over another are made.

Selecting measures and analyzing and interpreting data across the period of early childhood is challenging because, as in other domains, adequate measures are not available to comprehensively assess language development continuously from birth through age six. (The Preschool Language Scales may be used from birth through age six, but as described below in the section *Recommendations for Instruments*, the measure does not provide a sufficiently comprehensive assessment across key areas of language development during this period.) Both norm-based and criterion-referenced measures that can assess children's progress across infancy and early childhood must be developed with existing longitudinal data. Criterion-based measures are needed with benchmarks for monitoring progress, and these should be validated against standardized tests. Normative longitudinal data being used to develop instruments for identifying children with early signs of Specific Language Impairment (SLI) may be used for this purpose. Ultimately, the development of aligned criterion and norm-based measures requires new normative data that specifies age-referenced developmental milestones, benchmarks, and developmental trajectories for diverse populations.

Mathematics

Most standard measures in early mathematics are discrepant with recent scientific knowledge. That is, they are limited to narrow sets of low-level knowledge and thus do not adequately cover the knowledge,

skills, and processes that recent research indicates should be assessed. In addition, existing standard tools do not assess gradual transitions that occur between early preschool competencies and more formal mathematics of the early grades. Several promising instruments currently being developed would be appropriate for focused studies of early mathematics, but most still lack comprehensive coverage. Thus, a priority is to generate standard assessment procedures, perhaps by using combinations of items from newly developed measures, to cover all components of early mathematics. Especially needed are shorter, efficient versions for use in large-scale research. Most measures are norm-based. To evaluate whether instructional strategies promote children's progress and target essential precursor knowledge and skills, criterion-referenced measures with age-based benchmarks and dynamic assessments are needed based on developmental sequences and learning trajectories.

It is not known whether age-based norms would necessarily be accurate or meaningful in the area of mathematics because environmental effects, such as the early childhood curriculum, would be expected to modify these. As for the other domains of development, research is needed on the combination of individual, social, cultural, instructional, and neurobiological factors that impede or constrain the development of mathematical knowledge, skills, and processes at different ages. Studies should completely document early precursors of formal mathematical knowledge and skill, and clarify relations among constructs. This documentation is important for knowing which areas are essential to include in large-scale standard assessments. Such studies are also needed to answer questions such as: How do informal understandings of mathematics function as precursors to the development of formal knowledge?

Social-Emotional Development

Most standard measures of social-emotional development consist of teacher and parent reports. The report measures available tend to focus on children's negative behaviors (e.g., the Achenbach CBCL). The combination of positive and negative behaviors that make the optimal assessment is not known. Additional measurement development is needed to devise an efficient instrument that incorporates both positive and negative behaviors and that would be feasible for use in large-scale research.

In addition to further developing teacher and parent report measures, standard direct assessments must be developed using unstructured and structured observational methods. In-depth assessments should be designed for focused, small-scale studies and more efficient versions for large-scale research. Unstructured and structured observational measures should be theoretically based to ensure construct validity and a relatively small set of dependent variables that yield meaningful data that can be easily analyzed and interpreted. Protocols for collecting and coding observational assessments must be developed that measure social competency in the context of interactions with adults and with peers during social-play and other dyadic and group contexts.

Though context dictates the meaning of observed behaviors (e.g., aggressiveness), or whether children have the opportunity to display particular competencies, contextual information typically is not considered in the coding and analysis of behavioral observations. In addition, indexing behaviors relative to the local group yields data that are sensitive to intervention. Thus, instruments should be developed that incorporate observing, coding, and indexing behaviors according to theoretically based and well-specified context parameters. (One practical and ethical consideration, however, is that indexing procedures lead to the identification of individual children and require active consent, which can be accomplished but makes research more difficult and costly.)

In addition to developing new observational measures, research is needed to expand and/or improve observational components of existing measures (e.g., Leiter Examiner Ratings, and Bayley Scales of Infant Development). The NCAST (Barnard, 1994; Sumner & Speitz, 1994) is a highly efficient

assessment of social competency for use from birth through age three. It consists mostly of parent or caregiver measures, but also contains a set of teaching tasks for direct child assessment. These tasks take ten minutes to administer, and are appropriate for culturally and ethnically diverse groups. Additional research is needed to improve reliability and internal consistency, especially for the child measures. Methods used to assess adult-child interaction in the context of research on language development and joint attention may be developed to allow assessments of social development.

Most measures focus on adult-child interaction, yet peer interactions may be as important as interactions with adults in promoting learning and development, especially because they afford interactions, such as conflict, that tend to occur less often with adults. Standard protocols should be developed for observing, coding, and analyzing interactions with peers, and studies conducted to determine how peer interactions affect learning within group contexts. Though much research has been conducted on social problem-solving, social cognition, social meta-cognition, and the related area of "theory of mind," measures either have not been fully developed or validated. Some researchers have begun to adapt paradigms and measures used for measuring peer interaction and social problem-solving with older children for use with younger children, but much more work needs to be done. Measures are available for assessing the quality of peer relationships in kindergarten and the early grades, and these have been used in research showing the importance of high quality peer relationships to creating positive learning environments and to children's school achievement. However, a priority is to develop measures useful for assessing the quality of peer relationships and their effects on learning during the period before kindergarten.

Regulation of Attention, Behavior, and Emotion

Standard measures with sound psychometric properties do not exist to assess the regulation of attention, behavior, and emotion during the period of early childhood, though several promising measures may be further developed. Studies are needed, especially interventions that include behavioral tasks combined with neurobiological techniques (e.g., fMRI, EEG, etc.), to clarify relations among constructs. For example, relations between intrinsic motivation and task persistence and between sustained attention and persistence are not known. Research is needed to determine the combination of processes that allow the performance of complex responses involved in learning new behaviors and skills, such as inhibition of a pre-potent response in order to perform a new behavior, the capacity to hold a rule in mind and simultaneously execute a new action, and the ability to approach tasks in a planful and organized manner without being distracted by irrelevant stimuli.

Disagreement exists about whether some components of self-regulation are part of temperament, and thus individual, stable characteristics that are not malleable or sensitive to intervention. Available instruments reflect this conceptual confusion. That is, some instruments label items associated with behavior regulation as temperament, whereas others label them as social behaviors. Research is needed to clarify relations among these constructs and to determine relative malleability among subcomponents of regulatory behavior. Studies must identify the activities, environments, and styles of interaction and instruction that promote the display and development of regulatory skills depending on individual differences in temperament. Particular environments afford or prevent display of self-regulation and should be considered when using observational assessments or report measures.

Research has been conducted on generalized engagement, social engagement, and content-specific engagement in pre-academic areas such as literacy and mathematics, but standard measures of engagement do not exist. Though not sufficient for change, engagement is probably necessary for learning, and may be considered a benchmark, or precondition and thus should be assessed as a marker of program performance and effects. Changes in different types of engagement should be measured, as well

as the conditions that promote or impede it. More generally, research is needed to determine how engagement is affected by a combination of individual, environmental, social, neurobiological, and instructional factors.

E. RECOMMENDATIONS FOR INSTRUMENTS

Researchers categorized instruments into one of three tiers according to whether or not they are well-standardized, promising experimental measures, or appropriate for use in large-scale research relating to early childhood interventions or programs. Additionally, researchers identified a number of measures that enjoy popularity (and in many ways resemble those in Tier 1) but that are not recommended for continued use, especially in early studies of intervention. In these cases, researchers identified particular concerns that tended to be unique for each identified measure. Of particular interest were measures that could be used or developed as core instruments and that would allow the combining of data across studies to address larger sets of research questions, compare results, and conduct meta-analyses. The list is not intended to be exhaustive, but will provide a starting point for further discussion and for developing a more comprehensive compendium. This section summarizes comments participants made during discussion, but does not give a full description of each instrument. For published measures, a reference is provided for complete information. For experimental unpublished measures, the researchers developing the measure are indicated followed by a key reference, if available.

- **TIER 1** – required for inter-study comparison; published and widely used; well-normed; valid and reliable; sensitive to instruction or intervention; typically require minimal training and not labor-intensive; may include measures for more in-depth assessment.

- **TIER 2** – less frequently used; standardized and generally psychometrically sound; could be useful depending on the context, but less recommended; some observational measures may require high levels of training and/or be more labor-intensive; includes measures useful for more in-depth assessment within domains or for focused intervention.

- **TIER 3** – experimental, not published; considered promising; theoretically driven (typically based on new conceptual models); currently lack norms and psychometric validation; some observational measures may require high levels of training and/or be more labor-intensive; includes measures useful for more in-depth assessment within domains or for focused intervention.

- **NOT RECOMMENDED** – typically widely available and published; generally have established psychometric properties; not typically observational; tend to have some other characteristics that make the measure less desirable for research on the effectiveness of comprehensive early childhood programs.

Language Measures

LANGUAGE: TIER 1

Bayley Scales of Infant Development (BSID), Second Edition. The Bayley is widely accepted with documented, sound, psychometric properties, but an assessment that aligns better with goals for early childhood interventions is needed. When using the measure, investigators should distinguish between verbal and nonverbal items. A shortened version is in development for use in large-scale

national surveys, but concern was expressed that content and construct validity may be sacrificed for reliability, predictive validity, and ease of administration (Bayley, 1969;1993).

Peabody Picture Vocabulary Test (PPVT). This widely used measure of receptive vocabulary correlates with quality of child-care experience. Greater care should be taken to use the PPVT as it was intended. Interpretations of results obtained with the measure often go beyond the data without strict consideration of what the PPVT is known to assess and what it does not. Disadvantages of the measure are that administration time is long and repeated testing can dramatically affect a child's score (Dunn & Dunn, 1997).

Expressive One-Word Picture Vocabulary Test (EOWPVT). One caveat for use in some research designs is that improvements in scores often seem to be the result of improved test-taking ability rather than expressive vocabulary development; thus, it would be most useful in a randomized trial (Brownell, 2000).

Preschool Language Scale. (Comments refer to the 3rd edition). This standardized instrument of expressive and receptive language for use from birth to age seven is sensitive to intervention, and relatively easy to administer reliably. A learning effect has been demonstrated with repeated administrations. Researchers recommend combining this measure with others or using a more comprehensive measure of language because it does not sufficiently cover phonology, morphology, or the structural aspects of language. The recently published fourth edition does provide some coverage. Versions are available in Spanish and other languages (Zimmerman, Steiner, & Pond, 2002).

Reynell Developmental Language Scales. This measure for use from 15 months through age six years is considered a gold standard in measuring language comprehension and expressive language, especially the identification of slow development in these areas. A limitation is that it does not contain phonology (Reynell & Gruber, 1990).

Clinical Evaluation of Language Fundamentals: Preschool. This well-regarded and widely used, standardized measure uses a naturalistic administration procedure to test children's receptive and expressive language skills. In some circumstances, it may be preferable to the Preschool Language Scale (described earlier), especially if a shorter test is needed that is easier to administer (Wiig, Secord, & Semel, 1992).

LANGUAGE: TIER 2

Communication & Symbolic Behavior Scales (CSBS). This prelinguistic measure of joint attention assesses a known precursor or condition of language development. Though some questioned the psychometric soundness of this measure, others have found the CSBS to be a better predictor of later academic competencies than the BSID, and consider it useful. Data show the scales are useful for setting intervention goals for some children but more data are needed. A concern was that the structured format could limit its predictive value (Whetherby & Prizant, 1993).

LANGUAGE: TIER 3

Evoked response potentials (ERP). ERP measures of speech sound discrimination (e.g., speech versus non-speech) during infancy predict later language and reading abilities. This method is being used for infants from birth to age 18 months as part of large-scale screening. It has also been used with three- and four-year-olds. Standardization and more psychometric data are needed. Dennis Molfese and Victoria Molfese at the University of Louisville, and Paul Yoder at Vanderbilt University are conducting

this work. (See Molfese, Burger-Judisch, Gill, Golinkoff, & Hirsh-Pasek, 1996 for a description of how the procedure is used with adults.)

Phonological Precision Measure. This measure of articulatory clarity provides a method for examining the internal representation of the sound system. It involves puppet play, which children find enjoyable, and takes little time to administer. The measure is based on a theoretical model, which specifies that the quality of phonological representation underlies the observed connection between poor phonological awareness and poor reading skills. Data showing links between speech disorders and reading disorders are typically cited as support for the theory; however the hypothesis requires further testing with better designs that control, for example, the confounding of speech development and language development. The measure predicts decoding and reading achievement (Elbro, 1996, 1998).

Frog Stories. Picture books are used to elicit original stories and to prompt story re-telling, with the pictures constraining the story to be told. Children like the task and researchers consider it useful because it yields multiple quantitative measures of expressive language, information on language complexity, narrative structure, and emotional valence. Collection, coding, scoring, and analysis of narrative data is labor intensive and requires commitment; however, it may be the only measure for the period of early childhood that covers morphology, syntax, and phonology in one assessment. Informal protocols for coding and scoring data exist but standard protocols for administering, coding, and analyzing data are needed. Guidelines for scoring and coding in Spanish and English are being used in a larger study to develop an instrument to assess Spanish-English bilingual children for language impairment (Aquilies Iglesias at Temple University). Data from both telling and re-telling have been collected on more than 3,000 narrative coding schemes applied, and basic measures obtained such as mean length of utterance (MLU), number of words, number of different words, and words per minute. The stories also are being used in studies of Spanish-English bilingual preschoolers that include measures of phonology (Carol Hammer at Pennsylvania State University).

In other research, protocols and scoring procedures have been developed for use with third and fourth graders, but in one normative sample of third and fourth graders a significant percentage of children provided impoverished narratives, leading to concerns about the amount and accuracy of data that could be achieved with younger children. Researchers noted that even one-word responses labeling pictures could be useful for assessment; yet it is critical to use multiple methods to determine if only language capacity affects responses or if other characteristics, such as shyness, affect performance on this measure more than other methods of assessment. Researchers cautioned that investigators should become knowledgeable about the different coding systems that exist, but should select a single system and apply it consistently and reliably. A software program (SALT; Miller) may be used to generate scores for some variables automatically, but strict adherence to rules for transcription is required. (See Berman, Ruth A., & Slobin, D. I., in collaboration with Aksu-Koc, A. A. et al., 1994.)

Narrative Elicitation Task. This task adapts five stories from a narrative elicitation procedure developed by Shapiro & Hudson (1991). The task is easier and quicker than Frog Stories, discussed above. The data are promising, but much more work needs to be done to develop the task into a standard assessment. Susan Landry at the University of Texas-Houston Health Sciences Center developed the task for use in research on the effectiveness of early childhood education interventions.

Bilingual English Spanish Assessment (BESA). This new measure was designed to identify children with Specific Language Impairment (SLI), and therefore does not include every language construct. The semantics, phonology, and pragmatics subtests of this assessment would be useful for the general population; however, the morphosyntax items were selected to identify children with SLI. It would be useful to develop the morphosyntax subtest for broader use. Aquiles Iglesias at Temple University is developing norms for Latino children.

Syntax Comprehension Task. This receptive task for measuring comprehensive sentence comprehension is for use with four- to five-year-olds. Comprehension of complex sentences correlates with the complexity of teachers' sentences, indicating that ways of intentionally promoting use of complex syntax in the classroom is an instructional method that should be investigated. Standardization and psychometrics are needed. Janellen Huttenlocher at the University of Chicago developed this task.

Measure of African American English. This new measure, which takes about 45 minutes to administer, is being developed to distinguish children with SLI, and includes variants of the Nonsense Word ("wug") test. Charlena Seymour at University of Massachusetts is developing this measure.

LANGUAGE: NOT RECOMMENDED

MacArthur Communicative Development Inventory (CDI). This widely used, parent report measure of receptive language is normed and has documented psychometric properties. Most participants did not recommend the version in current widespread use, primarily because it lacks sensitivity to intervention, especially for diverse populations. Other researchers emphasized, however, that the CDI has been shown to vary with child-care quality, and thus is sensitive to environmental influences. A Spanish version is available. The CDI is currently being re-normed and used with large populations of children in studies of Head Start. These data are expected to be published soon and to lead to more conclusive evidence about the usefulness of the CDI for intervention research (Fenson, Dale, Reznick, Thal, Bates, Hartung, Pethick, & Reilly, 1993).

Literacy Measures

LITERACY: TIER 1

Woodcock-Johnson III Tests of Cognitive Abilities and Achievement. This comprehensive battery of cognitive and achievement tests was designed for individuals between the ages of two and 90. The sample included individuals representing a stratified random cross-section of the population with respect to age, sex, ethnicity, socioeconomic status, community size, and education level. The preschool sub-sample included 1,143 two- to five-year-olds from diverse geographic locations in the U.S who were not yet attending kindergarten.

Four subtests from the **Cognitive Abilities** battery assess **phonological skills** that include phonological sensitivity (the Sound Blending subtest and the Incomplete Words subtest), phonological memory (the Memory for Words subtest), and phonological access (the Rapid Picture Naming subtest). Internal consistency is very high for ages two to five years. Test-retest reliability obtained from intervals that range from one-to-10 years is moderately high. Correct Spanish responses are provided in the manual and have very high internal consistency for ages two through five years.

Two additional subtests from the **Achievement Battery** can be used to assess **print skills**: The Letter-Word Identification subtest and the Word Attack subtest (which requires that children correctly pronounce readable nonwords). The Letter-Word subtest was recommended also as a measure of decoding. Internal consistency and test-retest reliability for each subtest is high for ages two through five years. Concurrent validity data for these achievement subtests show significant but moderate correlations between the subtests and selected scores from other achievement batteries (McGrew & Woodcock, 2001; Woodcock et al., 2001).

Test of Early Reading Ability-3 (TERA-3). The TERA-3 measures early reading in children from three-and-a-half through eight-and-a-half years. Specifically, it consists of three subtests that assess the ability to attribute meaning to printed symbols (Meaning), knowledge of the alphabet and its functions (Alphabet), and understanding of the conventions of print (Conventions). Standardization for the TERA-3 began in early 1999 on a sample of 875 children from 22 states.

Reliability for the TERA-3 is moderate to high and internal consistency is high across all subtests. Validity scores, which are correlations between the TERA-3 and similar measures, range from .34 to .98. Its construct validity has been evaluated. The tool is especially valuable if an assessment is needed that focuses exclusively on key areas of print awareness. It is easy to administer, requires few materials, and takes approximately 15 to 45 minutes, depending on the child's age and ability. A software scoring system is available and provides classroom-wide and school-wide data, which can be used to establish local norms (Reid, Hresko, & Hammill, 2001).

Dynamic Indicators of Basic Early Learning Skills (DIBELS). This measure is appropriate for children in kindergarten or the early grades; it measures a range of literacy skills that are sensitive to intervention. Based on an IRT model, it contains randomly selected items that prevent learning effects with repeated administration and inaccurate results that can occur from "teaching to the test"(Good & Kaminski, 2002).

LITERACY: TIER 2

Get Ready to Read (Screening Tool). This 20-item screening tool covers pre-literacy constructs that include phonological awareness and print awareness. It contains research-based questions that are appropriate to administer during the year before kindergarten to determine whether children have the early literacy skills needed to become readers. Advantages are that it is quick and easy to administer. The measure is appropriate for use in impact studies, and may be appropriate for early identification of children for primary prevention; however, more psychometric work is needed. It is based on the research of Drs. Grover (Russ) Whitehurst and Christopher Lonigan. Members of the advisory panel included Drs. Jack Fletcher, Victoria Molfese, and Joseph Torgesen. (See National Center for Learning Disabilities, 2002.)

Woodcock Johnson III Tests of Cognitive Abilities and Achievement (WJ-III) (Sound Awareness Subtest). This subtest in the Achievement Battery can be used to assess phonological skills. It contains four brief subsections that assess phonological sensitivity (i.e., rhyming, deletion, substitution, and reversal) skills. Internal consistency for ages two through five years is high; however, participants indicated that it begins to be useful at age three. Psychometric data have not been documented for this particular subscale, which is needed for use independent from the larger battery (McGrew & Woodcock, 2001; Woodcock, McGrew, & Mather, 2001).

Woodcock Reading Mastery Tests-Revised (WRMT-R). This measure is designed to assess reading readiness and achievement from age five years through adulthood. Standardization occurred between 1983 and 1985, on 6,089 participants from 60 diverse geographic areas of the United States. Characteristics of the norming sample were selected to be representative of the U.S. population as reported in the 1980 U.S. Census. Form G includes four tests of reading achievement, two readiness tests, and a supplementary letter knowledge checklist. (Form H does not include the readiness tests or the supplement.) Basal and ceiling procedures require that children receive only a subset of the total number of items. This procedure makes the measure more time-efficient, but takes training to administer.

Within the achievement battery, there are two Basic Skills subtests, Word Identification and Word Attack, which assess children's pronunciation of words and non-words and take approximately 10

minutes to administer. Two readiness subtests are included: Visual-Auditory Learning and Letter Identification, which together take 15 minutes to administer. For the Visual-Auditory Learning subtest, children watch as icons are associated with words and must report the meaning of each icon and correctly label novel icon combinations. For the Letter Identification subtest children must label upper- and lowercase letters presented in a variety of print and font types. Finally, the Letter Checklist subtest requires oral labeling of the names and sounds of upper- and lowercase letters. (Two other achievement subtests in the Reading Comprehension cluster, Word Comprehension and Passage Comprehension, are not considered appropriate for young children who are not yet reading words.)

Internal consistency reliability and validity data on the individual subtests, clusters and total scores are available for selected grades and ages. However, no reliability or validity data are available for children in preschool or in kindergarten, accounting for its Tier 2 status. These data are needed on samples representative of the current U.S. population. (Woodcock, 1987)

Woodcock-Johnson III (General Knowledge Subtest). Though generally psychometrically sound, additional data are needed concerning construct and predictive validity. (Participants categorized this as Tier 2 for this reason.) For example, the subtest correlates with vocabulary measures, but research is needed to discriminate whether the measure uniquely predicts schooling effects and reading comprehension. Other potentially useful subtests include science, social studies and humanities for those seeking to assess early scientific and related basic information concepts (McGrew & Woodcock, 2001).

LITERACY: TIER 3

Preschool Comprehensive Test of Phonological Processes (Pre-CTOPP). This measure comprehensively assesses phonological processing, short-term and longer-term phonological memory, phonological sensitivity, phonological awareness and print awareness. It is designed for three- to five-year-olds from diverse backgrounds, and a Spanish version is in development. Good reliability has been demonstrated. It will be usable by fall 2002, but will not be standardized until winter of 2003. It would compete with the Woodcock-Johnson-III, and research is needed to compare the measures. Christopher Lonigan at Florida State University and colleagues are developing this measure (Lonigan et al., in preparation).

School-Home Early Language and Literacy Assessment (SHELL). This new instrument uses the narrative book "Snowy Day." One concern was that the theme may be difficult for southern children to understand, leading to discussion of the more general need to take material content into consideration when validating assessments with geographically and otherwise diverse populations of children. David Dickinson and Catherine Snow at Boston University are exploring the usefulness of this measure for early childhood intervention research.

Verbal memory and speech production. This measure developed by Gathercole and colleagues is sensitive to schooling effects and can be administered to children as young as three years of age; data are needed on construct and predictive validity (See Gathercole & Pickering, 2000, 2001).

Inventive spelling. Measures of inventive spelling were suggested for use in research with three- and four-year-olds, though they are typically used in research with kindergarteners and older children. Thus, existing measures may show floor effects and require further development. Scoring metrics are available.

Letter Identification/Naming Tasks. Some participants recommended using the entire alphabet rather than sampling letters whenever the goal is to assess the letter knowledge of an individual child. However, others recommend that sampling procedures be used because data show that the scores are

extremely reliable indicators of a child's letter knowledge. Sampling results in an accurate, cost-efficient, and useful measurement tool when a large number of individual children must be tested for program evaluation or research purposes. This judgment was based in part on reliability data obtained from testing children's knowledge of 15 letters in a free response format, and eight letters in a multiple choice format. Because there can be discrepancies between children's ability to name upper-case and lower-case letters, it could be useful to measure both; however, there is a high degree of consistency between the two.

LITERACY: NOT RECOMMENDED

Concepts of Print. Though used frequently, scores on this measure do not consistently predict literacy development and often lack unique predictive validity after accounting for other factors. Data patterns across studies indicate that the measure is a proxy for other unmeasured constructs that are more directly related to reading development. It probably measures constructs that are affected by the intervention but are not the precursors of reading development per se. Thus, when a study shows that an intervention improves concept of print scores, the results cannot be interpreted as showing that the intervention is developing early pre-reading skills that are essential to later reading achievement. Other measures should be used that assess development in areas that directly affect later reading ability (Clay, 1979).

Mathematics Measures

MATHEMATICS: TIER 1

None of the published instruments could be recommended for comprehensive assessment in mathematics. Most are general purpose and outdated instruments that are inconsistent with the current state of knowledge and with contemporary goals of early mathematics education.

MATHEMATICS: TIER 2

The Test of Early Mathematics Ability (TEMA-2). The TEMA-2 is a widely used individual assessment measure appropriate for ages three through nine. It is based on current theory and research and is grounded in normative data. It assesses both informal and formal knowledge, and is a good predictor of mathematics achievement. Coverage for preschoolers includes early numerical concepts and procedures (e.g., counting, enumeration, perception of more, concrete addition, and other aspects of number). The examiner has the opportunity, in a separate session, to identify the child's thought processes, estimate learning potential, and select appropriate educational interventions. Though the TEMA-2 gives useful measures of number development, it does not cover other aspects of mathematics, such as shape and spatial relations. A Chinese translation is available but without norms (Ginsburg and Baroody, 1990; Ginsburg, 1990). The third edition, TEMA-3, will be published in 2003. It will have parallel forms and more extensive norms (Ginsburg and Baroody, 2003; Ginsburg, 2003).

MATHEMATICS: TIER 3

Building Blocks Math Assessment. This measure covers ages three to seven years and emphasizes numeracy and geometric thinking. Numeracy items include counting (from verbal production to advanced counting strategies), subitizing (quick recognition of number), comparing and ordering numbers, nonverbal and verbal arithmetic, and general quantitative reasoning. The geometry items include shape recognition and naming, attributes of shapes, construction of shapes, comparing shapes (congruence), shape composition and decomposition, and spatial reasoning. Measurement and patterning are also assessed. A notable advantage of the assessment is its frequent use of constructed responses

items. This characteristic, and the scoring scheme, allow the assessment of mathematical processes and strategies as well as accuracy. Another unique feature is that it builds upon and assesses children's learning trajectories in each area. The assessment is administered individually, using a few common objects and manipulatives. It has been administered to hundreds of children, and its predictive validity is currently being assessed. Doug Clements and Julie Samara at the State University of New York at Buffalo are developing this measure (see www.gse.buffalo.edu/org/buildingblocks/).

Child Math Assessment (CMA). This measure under development for use with three- to six-year-olds has not been standardized, but is considered promising. It contains 56 items, administered in two 20-minute sessions. The CMA assesses change across time and is sensitive to intervention. The domains covered include: counting, knowledge of ordinal number, informal arithmetic, pattern knowledge, geometric knowledge and reasoning (e.g., analysis of shapes and spatial location), and measurement (e.g., weight, length, capacity; also nonstandard units). Responses are scored either correct or incorrect, with classroom observations to code behavior in math activities. An advantage of the CMA is that it assesses problem-solving strategies; therefore improvements in strategy can be observed even if a problem is scored incorrect at the beginning and end. These data can be useful for instructional purposes. A recommendation was the development of a standard version for use in large-scale research that covers all essential constructs, but contains fewer items from each subscale. It has been used with about 700 children in the United States, China, and Japan. Prentice Starkey and colleagues are developing this measure at the University of California, Berkeley.

MATHEMATICS: NOT RECOMMENDED

Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI). Though widely used with psychometric characteristics of a TIER 1 measure, some experts reported that the WPPSI yields only a superficial assessment of mathematics abilities and thus should not be used to evaluate the effectiveness of early childhood education programs (Wechsler, 1989).

Woodcock Johnson-Revised Tests of Achievement (Mathematics Subtest). The Woodcock-Johnson is administered using standard procedures, and can be used starting at age four years and with children from families with lower incomes (though floor effects have been reported). It is IRT-scored, with a developmental metric. Though it has characteristics of a TIER 1 measure, it was considered only minimally useful. The measure is not based on current research on the development of mathematical thinking. It does not adequately assess developmental progress that occurs between very low-level abilities and more advanced, formal knowledge. Thus, for the period of early childhood, it results in the assessment of very narrow, low-level content that includes numbers of three or less. Some participants believed that, if used correctly, it would be preferable to using no measure of early mathematics. However, researchers cautioned against a frequent misuse of the scale that involves selecting a few items that assess only low-level knowledge, which often occurs in large-scale research and compromises content validity (McGrew & Woodcock, 2001).

Bracken Basic Concept Scales-Revised. This general measure of school readiness was not designed to assess mathematics separately; however, it is one of the few standard measures with psychometric data available for assessing mathematics. Scores for correct responses predict mathematics achievement in first grade, and has norms for children as young as two-and-a-half years. However, the measure has poor content validity (some domains are missing) and only correctness is scored. A Spanish translation is available, but lacks norms. As with the Woodcock Johnson (Math subtest), it has extremely limited coverage and would be recommended only because other standard options do not exist (Bracken, 1998).

General Cognition Measures

GENERAL COGNITION: TIER 1

The three achievement subscales of the Woodcock-Johnson Revised were recommended (McGrew & Woodcock, 2001), as well as the Stanford-Binet Test of Intelligence (Thorndike, Hagen, & Sattler, 1986) and the WPPSI (Wechsler, 1989), which has an acceptable short version.

Researchers recommended that all studies include a measure of general intellectual ability (IQ) that is sensitive to intervention and can serve as a control needed to show the uniqueness and specificity of an intervention. Though short forms often do not meet the psychometric standards of the full-length versions, they may be considered under the time and cost constraints of large-scale research. Nonverbal measures may be useful for language minority groups, but researchers cautioned that both verbal and nonverbal measures require language proficiency.

GENERAL COGNITION: TIER 2

No general cognition measures were recommended in this category.

GENERAL COGNITION: TIER 3

No general cognition measures were recommended in this category.

GENERAL COGNITION: NOT RECOMMENDED

Kaufman Assessment Battery for Children (K-ABC). This measure was generally not recommended though widely used and standardized with documented psychometrics. The measure is based on outdated theories from the 1960s. Some subtests might be useful if theoretically motivated, but the macro-level scores are not recommended. (Kaufman & Kaufman, 1983)

McCarthy Scales of Children's Ability. This measure assesses cognitive ability and gross and fine motor skills from age two-and-a-half to eight-and-a-half years. It has not been renormed since 1972 (McCarthy, 1972).

Social-Emotional Development Measures

SOCIAL-EMOTIONAL DEVELOPMENT: TIER 1

Bayley Infant Behavior Record. The measure is intended to supplement information obtained from the BSID. Although general and widely used, there are a number of concerns about the appropriateness of the measure for most constructs within this domain. Its utility for measuring progress in the context of intervention research was questioned because wide variations in performance observed early in development may obscure later developmental changes that occur, especially across shorter periods of time (Bayley, 1993).

SOCIAL-EMOTIONAL DEVELOPMENT: TIER 2

Social Competence Behavioral Evaluation (SCBE). This teacher-report measure has a long version, a short 30-item version, and a parent version. The content of the measure is recommended. It includes a Social Competence scale that is well-differentiated as well as an Aggression-Anger scale and an Anxiety-Withdrawal scale. Responses of experienced teachers tend to be distributed differently from

inexperienced teachers, an issue to take into account generally when using data from teacher reports. The standardization samples are not large but considered adequate. The measure was not developed strictly for clinical use, though it correlates with the Child Behavior Checklist. The item content allows evaluations to be completed by anyone knowing the child well (La Frenier & Dumas, 1995).

Infant Toddler Social Emotional Assessment (ITSEA). This dyadic measure, designed for use with children from birth through age three, assesses how infants relate to others. Constructs covered include regulation, attachment, withdrawal, social competency, and positive and negative affect. The content is recommended and scores predict cognitive development as well as behavioral problems. Though not experimental, more data are needed on psychometrics and usefulness to justify recommending as a standard measure for large-scale research. Assessments take 30 minutes; parent-report and short versions are available. Alice Carter and colleagues at the University of Massachusetts developed the measure (For psychometric data, see Carter, Briggs-Gowan, Jones, & Little, in press).

Attachment Q-Sort. This measure for use with children from infancy through 36 months presumably examines the quality of the relationship between the child and primary caregiver. The extent to which the measure assesses individual characteristics of the child is not clear. The measure would be recommended for use in studies of home-based programs and interventions, but some participants believed that more work is needed to demonstrate its validity and usefulness with preschool teachers. The potential culture-specificity of the standard criteria is one source of concern. Q-Sort scores predict later behavior problems, but the reason is not clear (see Waters, Vaughn, Posada, & Kondo-Ikemura, 1995). A short version of this assessment has been adapted for use in the U.S. Department of Education ECLS-B study (http://www.nces.ed.gov/ecls/); however, additional work is needed to demonstrate its psychometric properties.

Behavioral Assessment System for Children (BASC). This measure is used frequently in schools, especially in early studies of ADHD. It assesses adaptability, leadership, social skills, attention, and learning. A serious concern about the scientific validity of the scale was raised: After the empirically based items were developed, clinicians were allowed to add items believed to have clinical significance. Thus, it appears to over-identify active children as having clinical levels of ADHD (Reynolds & Kamphaus, 1998).

Social Skills Rating System (SSRS). Items for this measure are abstracted from the CBCL. Recommendations for the measure were mixed and constrained because reported norms contain unusual patterns and children are placed into categories that provide minimal and conflicting information. The total score has the most predictive power (for example, the total social score observed at 54 months of age predicts cognitive and academic scores), but most recommended that scores for the four subscales not be used separately. Others reported that the social skills scale could be recommended, but not the academic competence scale or the behavioral scale, which contains only 10 items. Criteria for categorizing children on the basis of the reported norms is poorly justified; raw scores are recommended. With improved scoring procedures and a stronger standardization sample that includes children from families with lower incomes, the SSRS could be a more highly recommended measure. (Gresham & Elliott, 1990) (This measure was adapted for use in the U.S. Department of Education Early Childhood Longitudinal Study-Kindergarten Cohort; http://www.nces.ed.gov/ecls/)

SOCIAL-EMOTIONAL DEVELOPMENT: TIER 3

Student-Teacher Relationships Scale (STRS). This promising teacher self-report measure assesses the quality of a teacher's relationship with a student in the areas of conflict, closeness, and dependency and classifies relationships between teachers and students as dependent, improved, or secure. It covers teacher-child dynamics, teachers' decisions about the child's school career, and the child's

future school adjustment. The measure blends theory on child-adult attachment with research on the importance of early school experiences in determining the trajectories of children's school progress. It is intended to identify student-teacher relationships that could benefit from intervention and support. The STRS can be used separately or with the Student and Relationship Support intervention program. Scores are moderate predictors of school success. Additional standardization and psychometric work is needed (Pianta, 2001).

Minnesota Preschool Affect Checklist. This promising observation measure yields a useful profile of the child, and can indicate strengths, problems, and risk factors in the following areas: positive, negative, and inappropriate affect in the context of peer interaction (e.g., expression of empathy); positive and negative engagement with the environment; response to frustration; and peer skills. It has a lengthy observation checklist that produces reliable data with trained laypersons. Though the measure has been used in intervention effectiveness studies, participants indicated that more psychometric work is needed for inclusion as a standard measure in the context of early childhood intervention research (Sroufe, Schork, Motti, Lawroski & LaFreniere, 1984).

Emotion Knowledge Puppet Task. This individual child measure of emotion knowledge is administered as a game using puppets with a standard protocol. It predicts peer adaptation, and has demonstrated reliability across nine months. Scores are more reliable than peer nomination. Additional standardization and psychometric work is needed. Susanne Denham at the George Mason University is developing the measure.

Peer Interaction Preschool Scale (PIPS). This scale developed by Myrna Shure measures problem solving in the context of "typical" social interaction dilemmas. Preliminary evidence suggests that scores differentiate children who have positive peer relationships from those who do not, and predict the quality of peer relationships in kindergarten. (Shure, 1992, 1996)

Regulation of Attention, Behavior, and Emotion Measures

REGULATION OF ATTENTION, BEHAVIOR, AND EMOTION: TIER 1

The group did not identify any Tier 1 measures for this construct.

REGULATION OF ATTENTION, BEHAVIOR, AND EMOTION: TIER 2

NEPSY. This developmental neuropsychological assessment for three- to 12-year-olds is a laboratory measure that was designed for use by school psychologists, neuropsychologists, and research psychologists to assess children with developmental disabilities and to develop effective intervention strategies. Attention/executive functions covered include: inhibition, self-regulation, monitoring, vigilance, selective and sustained attention, maintenance of response set, planning, flexibility in thinking, and figural fluency. Participants believed it could be useful in the context of early childhood intervention research, but would need further development (Korkman, Kirk, & Kemp, 1997). (The measure also detects strengths and subtle deficiencies in four other domains: language, sensory-motor functions, memory, and learning. Additional measurement development in these areas may also prove useful.)

REGULATION OF ATTENTION, BEHAVIOR, AND EMOTION: TIER 3

Cooper-Farran Behavioral Rating Scale. This 37-item scale measures self-regulation and independence. Teachers rate children on two subscales: interpersonal and work-related independence. Items addressing independence include staying on task independently and whether the child raises his or her hand. This measure uses a 1 through 7 rating, providing a range of scores that discussants viewed as a strength. Though designed primarily for kindergarten, it may be useful for younger children. Research has been conducted to adapt it for use with three-year-olds, but development is in the very early stages (Cooper & Farran, 1991).

Peg Task. This task measures effortful inhibitory control, regarded as one component of executive functioning, but the task also requires sustained attention. Successful performance emerges between ages three and four years, and scores predict later school achievement, as well as teacher ratings of social competence. The measure is not appropriate for tracking progress, but can be used in randomized designs to assess intervention impact. Strategies for promoting performance or other implications for intervention and instruction are not known. Use of the measure could prove useful in the development of integrated assessments or research to clarify relations among, for example, the development of language, pre-reading skills, and executive functions. Adele Diamond, who is at University of Massachusetts Medical School, developed this task.

Kochanska Battery. This measure of effortful inhibitory control is highly regarded, enjoyable for children, and measures change across time. Scores have been shown to vary according to parenting, but more data are needed to determine whether or not it is sensitive to intervention. Tasks of delayed gratification predict later school achievement and moral development. Piloting is needed to determine if coding different types of inhibitory control (e.g., attention, emotions, strategies) is feasible and useful for prediction and intervention. The battery has good documentation and is being used in two or three large-scale projects (e.g., to investigate early roots of violence). Being developed by Grazyna Kochanska at the University of Iowa in collaboration with Kathleen Murray (see Kochanska, Coy, and Murray, 2001; Kochanska, Murray, & Harlan, 2000).

Child Behavior Questionnaire (CBQ). This highly regarded questionnaire developed by Mary Rothbart is used primarily as a parent-report measure for assessing temperament. This research measure is not appropriate for assessing developmental progress or for applied use. It consists of 195 items that cover 15 dimensions. Participants suggested that sub-parts may be used and the measure could be shortened for more comprehensive assessment in large-scale studies. Shorter versions of the instrument have been developed in collaboration with Samuel Putnam. One version assesses all 15 dimensions using scales that are shorter than those found on the original CBQ. A second even shorter version provides scales for three broad dimensions of effortful control, negative affect, and surgency extraversion. A Spanish version developed by Dr. Carmen Gonzales is available. Substantial psychometric work is needed for both the full-length and short versions. In addition, research is needed to clarify relations among the constructs of effortful control, temperament, and social behavior. That is, effortful control is considered part of temperament in this scale, but in other scales with similar items it is considered a social behavior. Behavioral flexibility is one component of self-regulation believed to be important for scholastic success that may be embedded in the CBQ. Participants noted that the CBQ is written at a high literacy level. The reliability of the measure for use with lower socioeconomic samples was questioned unless adapted for populations with lower literacy levels. All CBQ versions may be obtained at http://darkwing.uoregon.edu/~maryroth/cbqdesc.html.

Leiter International Performance Scale-Revised (Sustained Attention Task). This measure is sometimes labeled as an assessment of persistence on challenging tasks, but is most widely regarded as a measure of sustained attention. Consisting of 50 to 60 items, this non-verbal assessment has national

norms, a standard score, requires little training, can be administered quickly, easily, and reliably, and is suitable for a broad age range. This task involves giving children, beginning at two to three years of age, a target (e.g., a flower) and then giving them an array of figures including the target. The child is asked to cross out as many of the target items as they find in a fixed amount of time. However, it has several limitations; 15 percent to 20 percent of the data are lost because children do not mark scoring sheets precisely. It shows practice effects and does not assess developmental change; thus, it should be used in intervention research with randomized designs to assess impact.

Leiter Examiner Ratings involve one hour of direct assessment that is used to generate a summary perception of the child. Parent and teacher forms are available and refer to children's behavior in broader contexts. Eight subscales consist of items believed to assess attention, organization, activity, sociability, feelings and mood, regulation, and sensory reactivity. All have unknown reliability and validity; however, it was recommended that this measure could be further developed and used to track change. Factor analyses and additional psychometric work are needed. Information is needed on the relation among children's performance on the sustained attention task, parent and teacher ratings of children's behavior, and standard observations of children's behavior in learning and social contexts (Roid & Miller, 1997).

Teacher Observation of Classroom Adaptation, Revised (TOCA-R). Three factors have been obtained for this teacher-report measure of child behavior: cognitive concentration, authority acceptance, and social contact. The cognitive concentration factor includes ratings of child characteristics such as self-reliance, ability to work alone, distractibility, eagerness to learn/curiosity, ability to complete assignments, sustained attention, ability to concentrate, and degree of task-oriented effort. The scale seems sensitive to intervention, has good reliability, and is useful for measuring developmental progress. Participants believed the TOCA-R is promising, but more development is needed that ensures appropriateness for measuring progress for three- and four-year-olds. The measure was developed and is typically used with children in kindergarten and early elementary school classrooms; however, it is currently being used with a Head Start sample in Baltimore. The scales are described in a technical report available from Kellam and colleagues at Johns Hopkins University and in a published study (Werthamer-Larsson, Kellam & Wheller, 1991). Additional information on the TOCA-R may be found at http://www.bpp.jhu.edu/publish/Manuals/TOCAmanual.htm.

Behavioral persistence. This measure of behavioral persistence in challenging situations is a structured observation of four-year-olds. John Love at Mathematica Policy Research, Inc. is developing the measure.

REGULATION OF ATTENTION, BEHAVIOR, AND EMOTION: NOT RECOMMENDED

Achenbach Child Behavior Checklist (CBCL). This frequently used, standardized report measure for children between ages one-and-a-half and five has sound psychometrics, and is widely regarded as a gold standard. However, it was generally not recommended for research in the context of early childhood programs because it was developed for clinical use and was believed to lack sensitivity for measuring typical development in a broader population of children. The checklist contains 112 items, many of which were believed to focus too exclusively on negative behaviors. Some participants expressed concern that the emphasis on negative behaviors often evokes teacher and parent resistance and could potentially affect the quality of data obtained. Others believe this concern is not supported across studies. Parents who judge an item does not apply can indicate it is not applicable. Developed by T.M. Achenbach, the newest revised version of measures and manuals is co-authored by L.A. Rescorla (Achenbach & Rescorla, 2000).

Because a single instrument does not efficiently or comprehensively cover negative and positive behavioral changes, participants recommended complementing the CBCL with assessments of pro-social behaviors to more comprehensively assess social-behavioral development. An additional concern was that the measure is written at a literacy level that may limit usefulness with low-income samples.

Strange Situation. Among other concerns about construct validity, cultural sensitivity, and practical usefulness, the amount of time and training required for administration for this laboratory assessment makes it undesirable for use in large-scale studies (Ainsworth, Blehar, Waters, & Wall, 1978).

Day/night Stroop. This task used to assess inhibitory control (and other constructs associated with executive functions) has been associated with early mathematics and reading ability in typically developing children and children with head injuries. Concerns were raised that the measure shows skewed distributions with young children and is not sensitive to intervention. More research is needed to determine whether this task has potential for measuring processes during the preschool years that are desired targets of early childhood interventions (Gerstad, Hong & Diamond, 1994).

PART TWO: DESIGNING A NATIONAL REPORTING SYSTEM FOR HEAD START

During the course of the meeting, participants were asked to provide guidance to the Head Start Bureau as it began the process of addressing its mandate to report on outcomes for all children in Head Start. Two assumptions guided discussion. First, regardless of how the Head Start Bureau responds to the mandate to assess each child's progress in language, literacy, and numeracy, programs will continue assessments underway and will use the directive as an opportunity to enhance these. Second, distinctions must be maintained between measures and data appropriate for program improvement and for accountability.

Nine design issues were proposed for consideration: the purposes of data collection and usage of data; outcome areas; measures; indicators of progress; aggregation of data; coordination of data with other Head Start data collection (e.g., impact study), confidentiality and associated protections; implementation; and professional development. Existing federal reporting systems and data collection on Head Start were reviewed to determine whether existing mechanisms or measures should be modified or replaced. Discussions focused primarily on design options, outcome areas, measures, conceptual frameworks, and components for the system.

A. Federal Reporting Systems Currently Underway

Two monitoring systems and two research studies are already in place or in progress. Head Start grantees are externally monitored once every three years for compliance to the Head Start Performance Standards using the **Program Review Instrument for Systems Monitoring (PRISM)**. One-third of programs are assessed once each year. Data are collected at different times of the year across sites. The **Program Information Report (PIR)** is an agency self-report that is approved by the administration's Office of Management and Budget and used to respond to the Government Performance and Results Act (GPRA) and public inquiries. The report includes the percentage of children and families receiving different types of services (including Immunization, Medical and Dental Health Services, Services for Children with Disabilities, Mental Health Services, Social Services for Families), the demographics of those served and descriptions of staff. Data are reported at both the grantee and delegate level. Local programs receive the report form the summer prior to the program year, submit by late spring, and data are available in fall. This mechanism can be modified extensively each year.

The **Head Start Family and Child Experiences Survey (FACES)** is a longitudinal study of the cognitive, social, emotional and physical development of Head Start children, the characteristics, well-being, and accomplishments of families, the observed quality of Head Start classrooms, and the characteristics, needs and opinions of Head Start teachers and other program staff. The first cohort was launched in 1997, with a nationally representative sample of 3,200 children from 40 home-based, classroom based, and family child-care Head Start programs. Children and their families were studied at program entry, after one or two years of participation and again at the end of the kindergarten year. A new national cohort began in fall 2000, with a sample of 2,800 children entering Head Start in 43 new Head Start programs (see Web site links and presentations above for more detail).

The **Head Start Impact Study** is a congressionally mandated longitudinal study of 5,000 to 6,000 three- and four-year-old children from a stratified, national sample of grantee and delegate agencies. Families with children applying for enrollment were randomly assigned to a treatment group to receive Head Start services or a comparison group. Constructs for measuring child outcomes for the two studies are defined similarly, with the Head Start Impact Study reflecting slight changes in instrumentation based on experience with FACES. Data collection began in the fall of 2002 and will continue through 2006. Children will be followed through spring of their first grade year. Data collected on individual children include areas related to school readiness, such as language and literacy, cognition and general knowledge,

social and emotional development, approaches to learning, physical well-being and motor development (see Web site links and presentations above for more detail).

B. Design Options

Participants were encouraged to consider a range of possible systems to achieve the goals of the reporting system, as articulated by federal staff. Their discussions produced five non-mutually-exclusive options.

(1) Expand FACES

Existing FACES measures could be modified to be more consistent with recent research on what outcomes to measure and how best to measure them. A matrix sampling approach could be used to collect data from children selected from Head Start programs. The data may or may not be longitudinal, but regardless could be aligned with on-going data collection mandated for other purposes (e.g., impact research). Two issues to decide would be whether to select children randomly and whether to sample some or all Head Start programs. To meet the administration's mandate of collecting data on each child in Head Start, Option 1 must be combined with either Option 3 or Option 5 below.

(2) Create a New System of National Data Collection

Creating a new system of national data collection would avoid having to decide which existing FACES measures to retain. All other elements of the system would be the same as Option 1. This system would replace FACES if it generated on going information needed for other purposes, such as GPRA reporting and local assessment. Option 2 must be combined with either Option 3 or Option 5 to meet the administration's mandate of collecting data on every child in Head Start.

(3) Use the Program Information Report (PIR) Mechanism and Data

A set of variables could be created that translates information from the grantee level into an annual, agency-level reporting system. For example, agencies could report annually on the percentage of program graduates who demonstrated acceptable progress (yet to be defined) in particular language and literacy competencies such as alphabet knowledge, phonemic awareness, early writing skills, vocabulary, as well as competencies in mathematics, social skills, and the other domains. Monitoring teams would, as part of their ongoing evaluation of systems, be able to verify this agency-reported data during on-site reviews. PIR variables could be modified periodically to focus on particular areas of interest. This option could be combined with Option 1 or Option 2 and would depend on strong ongoing support for Option 5.

(4) With Option 1 or Option 2, Add More-Intensive Assessments in a Particular Area Each Year, and Rotate Area of Intense Focus

In addition to the overall annual reporting on all outcome areas, intensive, in-depth information could be obtained on language and literacy once every three or four years, on mathematics once every three or four years, on social-emotional once every three or four years, and so on. This strategy has the advantages of providing broad information on areas of particular interest, permitting some large-scale data collection on measures developed through the Interagency Early Childhood Research Initiative, and providing opportunities to test measures for updating the ongoing annual system.

(5) With Option 1 or Option 2, Continue to Evaluate and Validate the Local Assessment Systems in each Head Start Program

The current mandate allows myriad local assessments, including tools that are locally developed. However, this system could be made more coherent and strengthened with a national validation process. The Head Start Bureau would first develop clear criteria that programs should use to select assessment tools (including reliability and validity, appropriateness for Head Start goals, etc.) and set standards for appropriate use. This would allow locally designed options while ensuring some commonality and use of scientifically validated tools, with accountability at the local level. An alternative is that, along with criteria and standards to support the selection and use of assessments, a list of empirically validated systems of approved assessments that align with curricula could be provided to grantees. Regardless of which approach is selected, training and technical assistance would support appropriate use, links to curriculum, data management, data analysis and interpretation, and data reporting to parents, staff, and external groups. Reviewers could use the criteria to judge the merit of applications and monitors could use standards to evaluate compliance. Both options allow for complete local decision-making and accountability. One suggestion was that public access to the criteria and standards would stimulate the development of measures appropriate for local use.

Some participants agreed that combining Options 1 or 2 with Option 5 leads to an especially strong system of assessment. A combination of national and local measurement and assessment approaches would link broad indicators for national reporting to classroom assessments teachers use to guide instruction. Such a system would have the capacity to produce converging data from national and local levels useful for determining which local assessment tools are effective and for evaluating the quality and effectiveness of Head Start systems.

C. Outcome Areas

The consensus was that appropriately monitoring outcomes of a comprehensive early childhood program requires comprehensive child assessment. The design of a national reporting system for accountability will dictate local program attention and priorities for instruction, and so from the piloting phase, it should cover all essential domains. If the system begins with a narrower focus, the likely result is a narrow focus at the program level. Given what is known about how children learn and develop, positive effects on outcomes are most likely if the system reflects current understanding of how developmental processes are connected across domains.

Many agreed that as a starting point the system should include the five broad dimensions of children's early development and learning outlined in the National Education Goals Panel (Kagan et al., 1995): language and literacy, cognition and mathematics, social-emotional development, approaches to learning, and physical development and health. These areas include important dimensions of children's development that are vital for success in school, reflect the general goals of Head Start programs, can be precisely defined and reported to yield profiles of children's development, and allow analysis of how development in one area affects the others. Such comprehensive coverage makes possible a more complex analysis of strengths and weaknesses useful for tailoring training and technical assistance.

A strong recommendation was that the data be able to show the amount of change and the conditions under which change does or does not occur. Information is needed, for example, on the demographics and other individual characteristics of children, families (e.g., parenting and home environment), neighborhoods, communities, classrooms, teachers, and programs. This broader context for reporting and interpreting the data is critical for understanding the circumstances that support or impede children's progress, including those that are and are not under the control of Head Start, and thus is useful for allocating training, technical support, and other resources.

D. Measures

Language and Literacy. Experts in language and literacy development presented a working model premised on the fact that key developmental milestones in language and literacy established through research are also useful for assessment and intervention in applied settings (see Appendix B). The model outlines key developmental periods of early child language and literacy, and within each period, describes what all normally developing children tacitly "know" (to be used as "outcome" goals) and age-appropriate means of assessment. Administration of assessments would be distributed across multiple sources (e.g., teacher, parent, specialist) and different components of the data collection would be useful at both national and local levels. Though the model is a "work in progress" and is focused on language and literacy, it could be a framework for guiding the selection and use of assessments in early mathematics, social and emotional development, and other crucial domains to use for national reporting and program improvement at national and local levels.

Key features of the model for monitoring and individual assessment include:

- Is grounded in scientifically validated facts of biological growth and of language development for typically developing populations with diverse socioeconomic and linguistic backgrounds
- Sets benchmarks that are universally agreed upon by developmental scientists
- Includes assessments that cover a fuller range of language and literacy constructs
- Goes beyond the mandate to include monitoring of language and pre-reading from birth to three years, to allow evaluation of interventions designed to prevent disparities in development already evident by age three years
- Offers ways for parents and teachers to better understand development and judge whether programs and individual children are moving towards developmental and educational goals

Key features especially useful for individual assessment include:

- Allows testing of every child
- Includes age-appropriate assessments informed by multiple sources
- Yields information useful to teachers for setting goals, generating strategies, planning assessments that cover essential areas, and learning whether strategies were successful
- Aligns closely with what children should know and have developed at particular points in development
- Includes child responses that are easily understood by parents and teachers, supporting easy and accurate completion
- Minimizes the number of specialists needed to conduct assessments because staff and speech-language therapists presently on-site could administer them (if appropriate funds were available to ensure the appropriate type and number of specialists at each site)
- Has potential to be used with all children including those with language disabilities and English Language Learners
- Includes measures that complement existing data collection (e.g., FACES and Head Start impact studies)

Mathematics. A complete assessment strategy in mathematics would consist of: 1) general indicators of progress, 2) in-depth assessment of specific knowledge and competencies, and 3) probes for teachers to assess children's concepts, strategies, and cognitive processes that underlie correct and incorrect responses. General indicators would be predictive of math achievement measured with nationally normed achievement tests for kindergarten or first grade. Specific areas to assess include number (counting and

counting strategies), subitizing (instant recognition of the numerosity of small sets), operations, shape (naming, attributes, construction, and composition/decomposition), spatial relations, measurement, and patterns. Assessments should evaluate formal, informal, and metacognitive (language, expression, self-correction) knowledge. The systematic use of probes should be encouraged to assess the level and type of reasoning that underlies children's responses in order to guide instruction. For example, standard protocols for administering tasks, conducting interviews, and making observations of children's problem-solving are modeled in a supplement to TEMA-2 (Ginsburg, 1990) that also includes recommendations for educational interventions. This assessment strategy goes beyond the narrow concept of numeracy to be consistent with current scientific evidence on the range of knowledge and competencies foundational to early mathematical development.

However, the TEMA-2 and other available instruments include only a subset of the areas needed for a comprehensive assessment. Therefore, a recommended solution was to develop and fully validate new measures for sufficient coverage, perhaps building from existing items across available instruments (e.g., items developed by Ginsburg and Baroody that assess number and operations; items developed by Clements and Samara to assess shape and spatial relations; items to assess pattern and other aspects of mathematics developed by Starkey, and items from other assessments such as the Work Sampling system).

The recommendation was made that Head Start experiment systematically with different types of assessments to generate new information on how best to evaluate and report on children's progress, while collecting a common set of data for national reporting. The data collected would be useful for reporting on child outcomes and informing teaching strategies. For example, in one condition, the TEMA-3 could be administered, which focuses primarily on number and operations. A second condition could implement a new test developed using items from FACES, the Early Childhood Longitudinal Study (Birth and Kindergarten cohorts), TEMA-3 and other sources. Additional conditions could test the value of adding components, such as use of technology to collect data for teachers to use to guide instruction, to either the TEMA-3 or the newly developed instrument. The conditions could be compared on a set of key factors, such as ease of administration, predictive validity in relation to desired outcomes, and so on. Dynamic and local methods of assessments to inform instruction would involve simple modifications of the measures described above that could be piloted. For example, teachers would be encouraged to examine children's informal and formal knowledge in all areas of mathematical development and probe underlying thought processes using standard methods. The utility of technology, such as hand-held computers, for evaluating the effectiveness of early childhood programs could be explored.

With respect to overall design options, a pre-post design was recommended in which teachers conduct assessments at the beginning and end of the academic year, and monitors trained to evaluate individual Head Start programs would conduct at least one external assessment. Ideally comparisons would be made between exemplary and non-exemplary programs. Data on the structure and stability of inter-correlations among different types of mathematical knowledge could be collected as well as information about how teachers can best use information from standard probes to promote children's progress.

Characteristics of Measurements and Assessments. Additional recommendations were made for selecting specific instruments and for approaches to measurement and assessment.

- Measures should be selected on the basis of recent scientific knowledge of developmental milestones and sequences, and indicate precursors of desired end points. This approach provides guidance for instruction by indicating what children should be expected to know at particular points in development and obliges selection of practices known to influence the development of specific precursor knowledge and skills.

- Often existing measures have not been normed on diverse populations, and Head Start children would be expected to show lower age-norms. This lack of validation with Head Start samples need not prevent the use of existing measures, though differences in performance between groups must be documented and measures should have developmental sensitivity for diverse populations. Floor effects must be avoided to ensure detection of whether or not progress is occurring. Measures currently normed on children younger than age three years or on different or more homogeneous samples might be considered because these measures may be sensitive for assessing where children actually are in their development.
- Measures for local use must be inexpensive to purchase and easy to implement, show growth over time, yield information parents can understand, and give profiles of strengths and weaknesses to guide instruction. They would ideally align with and be validated against indicators collected at the national level, and measures used in evaluation and impact research. In all of these respects, the Child Development Inventory (CDI) serves as a model for the development of new teacher and parent report measures for three to five-year-olds, though the measure itself is inappropriate for use beyond 24 months.
- Parent-report can be part of a gold standard system of assessment if it has demonstrated reliability and validity for the studied population; it should not however entirely replace direct assessments or assessments from multiple perspectives.
- Contextualized assessments with familiar materials, environments, and adults are critical for knowing how children use their knowledge and skills to function in daily environments. A combination of decontextualized and contextualized assessments is recommended because together they provide a more complete and accurate assessment of children's progress, strengths, and weaknesses.
- Methods of questioning children to determine the cognitive processes and strategies underlying accurate or inaccurate responses may be useful for guiding instruction. The TEMA-3 Assessment Probes provide an example of this approach. More research is needed to determine how best to obtain this information from very young children and how teachers should use it to improve instructional strategies.
- Some researchers recommended that portfolio/work-sampling methods designed to guide instruction should not be used for evaluation or reporting purposes because, though possible to translate into a reportable format, they do not yield data with psychometric properties suitable for aggregation and reporting at the local program or national level. If a national validation process were used to strengthen local assessments (Option 5 above), all local dynamic assessments would need to meet criteria for consistency with recent research and perhaps be validated against indicators selected for reporting at the national level.

E. Professional Development

A consensus was that defining and collecting data on outcomes at the local level offers an opportunity to educate practitioners about strategies for promoting children's learning and development that are grounded in scientific research. If poorly designed and executed, however, the system could encourage a focus on narrow skill sets and the use of rote instructional methods. Professional development would be crucial for doing accurate assessments and using the results to improve practice. It would strengthen the assessment system and promote understanding of the value of assessment for guiding and improving instruction. This understanding is essential for teachers as well as for the entire Head Start support network; therefore, all levels of the Head Start system should receive professional development including classroom staff, the training and technical assistance network, federal monitoring teams, and regional and central office staff. The need was emphasized for better coverage of assessment in all higher education programs, including Child Development Associates (CDA) and associate degree programs.

F. Conceptual Models and Components

A developmental conceptual model was presented in which periodic measures used for accountability would be instructionally meaningful. Scientific evidence on development would be used to identify periodic precursor measures that serve as benchmarks, that have predictive validity in relation to the desired long-term outcomes, and that align with local assessments that teachers use to guide instruction.

According to this model, any useful system of assessment and reporting would have the following characteristics:

- Well-defined end points or long-term outcomes that the program is purposefully designed to influence and that can be measured with reliability and validity (e.g., reading achievement at Grade 2)
- Periodic measures that assess known precursor knowledge and skills, which would be used both as benchmarks for targeting instruction or intervention and that provide data for accountability
- Instruments that yield metrics appropriate for assessing growth along the anticipated developmental trajectory
- Screening and monitoring using local dynamic assessments, combined with on-going and timely reporting of aligned indicators to determine which children, which programs, which regions, etc., are off-trajectory, by how much, and in which areas

The model uses the current mandate as a starting point, and thus begins from age three years, but it could be extended downward from birth for programs such as Early Head Start.

Several assumptions of the model were made explicit:

- It is possible to design and implement programs that gradually improve the average level of performance, with the most likely projected course being that smaller amounts of progress are made in the beginning, with larger gains over time.
- Teachers have timely information and program capacity needed to analyze whether their practices affect children's progress. Teachers have access to and are trained to use assessments at the local level that track whether or not children are progressing sufficiently, and have access to tailored technical assistance and professional development needed to select and implement the strategies. These strategies would be sensitive to a range of individual and contextual factors and avoid narrow and rote instructional methods.
- On-going local dynamic assessment systems used to monitor each child's progress for instructional purposes are aligned with and have validity in relation to the periodic measures collected at the national level. This alignment and validation process ensures that national indicators detect processes occurring at the local level, and that the information is useful for selecting strategies and allocating the appropriate technical assistance and resources needed to promote progress.
- The system has periodic assessments and outcomes that do not rely only on face validity concerning the content children should know at particular points in development (e.g., knowing exactly 10 letters) without evidence to support that belief. Instead, periodic assessments are selected using research on known precursors that make instructionally meaningful benchmarks for measuring children's progress toward desired end points that are measurable with standard instruments.

- The model is more consistent with criterion measures than normative measures because it focuses on the accomplishments needed to move from one point to another and on consistently raising performance.

Comments of the group were that the focus on growth and sequence, and the back-mapping strategy from desired outcomes through precursors is useful because it makes explicit the central question, what does the child entering kindergarten need to know and be able to do. It makes explicit that processes underlie outcomes and it does so in a way that is meaningful for teachers: it guides expectations, breaks down challenges, and points directly to clinical information that would be needed to reach each benchmark. At the same time, it illuminates the need for developmental research expertise to identify earlier forms and competencies, makes explicit the different forms and trajectories that are possible, and specifies strategies for influencing progress.

The general issue of who should collect data on child outcomes was raised, and specifically whether teachers should collect data used to hold programs accountable for performance. One comment was that it remains an untested assumption that data collected by teachers are necessarily invalid for accountability purposes. Teachers might collect valid data on child outcomes for both program improvement and accountability if: 1) data are useful to teachers and not only to researchers and evaluators, 2) teachers understand the value, and 3) capacity and resources are available to help teachers use the data to promote children's progress.

Two recommended features for the system were that it be able to: 1) accommodate modifications with advances in scientific knowledge and measurement, and 2) serve integrated research, technical assistance, and reporting functions. With respect to the latter feature, a national reporting system that includes matrix sampling and both national and local data collection could be used as a research base for reauthorization, with capacity to determine systematically which curricula and program services in use are effective and under which conditions. Hierarchical linear modeling was suggested as one quantitative methodology that could be used to study the combination of contextual and individual difference factors that produce particular developmental trajectories. The system would be useful for evaluating the quality, benefits, and liabilities of program decisions made at local and national levels, as well as indicate areas and strategies for improvement. Ideally, tracking of outcomes would be connected to other Head Start data collection (e.g., impact research).

G. Priorities for Research and Measurement Development

The models presented offer starting points for conceptualizing and implementing a national reporting system, as well as gold standards that may be met after gaps in basic knowledge and assessment tools are filled that currently limit the potential of such a system for program improvement and accountability. Complex, multi-level, longitudinal research using correlational and experimental designs were recommended for documenting developmental trajectories in each domain and determining how diverse and malleable growth trajectories can be, based upon a combination of individual, contextual, and instructional factors. Studies are needed that show amounts and rates of change expected if a program is fully and effectively implemented, and that isolate what the effective program components are. Few measurement tools are available for tracking change within a single area of development. The development of integrated assessments was recommended, but requires additional research on how developmental trajectories are related across domains, and how common cognitive, linguistic, social, affective, and neurobiological processes underlie development across multiple domains, such as early reading and mathematics. Ideally, developmental milestones from birth through age five years would be used to generate a seamless system of measurement to evaluate progress from birth into kindergarten.

IV. REFERENCES

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation.* Hillsdale, N.J.: Erlbaum.

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for ASEBA preschool forms and profiles.* Burlington, VT: University of Vermont, Department of Psychiatry.

Barnard, K. (1994). *NCAST Teaching Scale*. Seattle, WA: University of Washington, School of Nursing.

Bayley, N. (1969; 1993). *Bayley Scales of Infant Development, Second Edition: Manual.* San Antonio, TX: The Psychological Corporation.

Bracken, B. A. (1998). *Bracken Basic Concept Scale-Revised: Examiner's Manual.* San Antonio, TX: The Psychological Corporation.

Berman, Ruth A., & Slobin, D. I., in collaboration with Aksu-Koc, A. A. et al. (1994). *Relating Events in Narrative: A Crosslinguistic Developmental Study.* Hillsdale, New Jersey: Lawrence Erlbaum.

Brownell, R. (2000). *Expressive One-Word Picture Vocabulary Test Manual.* Novato, CA: Academic Therapy Publications.

Carter, A. S., Briggs-Gowan, M. J., Jones, S.M., & Little, T. (in press). The Infant-Toddler Social and Emotional Assessment (ITSEA): Factor structure, reliability and validity. *Journal of Abnormal Child Psychology*.

Clay, M. M. (1979). *Concepts About Print*. Exeter, NH: Heinemann.

Cooper, D. & Farran, D.C. (1991). *Cooper-Farran Behavioral Rating Scale*. Clinical Psychology Publishing Company, Inc.

Dunn, L., & Dunn, L. (1997). *Peabody Picture Vocabulary Test—Third Edition: Examiner's manual.* Circle Pines, MI: American Guidance System.

Elbro, C. (1996). Early linguistic abilities and reading development: A review and a hypothesis. *Reading and Writing: An Interdisciplinary Journal, 8*, 453-485.

Elbro, C. (1998). When reading is "readn" or "somthn." Distinctness of phonological representations of lexical items in normal and disabled readers. *Scandinavian Journal of Psychology, 39*, 149-153.

Fensen, L., Dale, P., Reznick, S., Thal, D., Bates, E., Hartung, J., Pethick, S., & Reilly, J. (1993). *MacArthur Communicative Development Inventories: User's Guide and Technical Manual*. San Diego, CA: Singular/Thomson Learning.

Gathercole, S.E. & Pickering, S.J. (2001). Working memory deficits in children with special educational needs. *British Journal of Special Education,* 28, 89-97.

Gathercole S.E. & Pickering, S.J. (2000). Working memory deficits in children with low achievements in the national curriculum at seven years. *British Journal of Educational Psychology,* 70, 177-194.

Gerstad, C.L., Hong, Y.J. & Diamond, A. (1994). The relationship between cognition and action: Performance of children 3-and-a-half to 7 years on a Stroop-like day-night task. *Cognition, 53,* 129-153.

Ginsburg, H.P., & Baroody, A. J. (1990). *The test of early mathematics ability: Second Edition*. Austin, TX: Pro Ed. (See also Ginsburg, H. P. [1990]. *Assessment probes and instructional activities: The test of early mathematics ability* [2nd ed.]. Austin, Texas: Pro Ed.)

Ginsburg, H.P., & Baroody, A. J. (2003). *The test of early mathematics ability: Third edition*. Austin, TX: Pro Ed. (See also Ginsburg, H. P. [2003]. *Assessment probes and instructional activities for the Test of Early Mathematics Ability-3*. Austin, TX: Pro Ed.)

Good, R.H. (2000). *Dynamic Indicators of Basic Early Learning Skills*. Eugene, OR: University of Oregon, Institute for the Development of Educational Achievement.

Gresham, F.M., & Elliott, S.N. (1990). *Social Skills Rating System: Manual*. Circle Pines, MN: American Guidance Service.

Kagan, S. K., Moore, E., & Bredekamp, S. (Ed.). (1995). *National Education Goals Panel. Goal 1 Technical Planning Group Report.*

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children: Administration and scoring manual*. Circle Pines, MN: American Guidance Service. (See also Kaufman, A. S., & Kaufman, N. L. [1983]. *Kaufman Assessment Battery for Children: Interpretive manual*. Circle Pines, MN: American Guidance Service.)

Kochanska, G., Coy, K. C., & Murray, K. T. (2001). The development of self-regulation in the first four years of life, *Child Development, 72*, 1091-1111.

Kochanska, G., Murray, K. T., & Harlan, E. T. (2000). Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology, 36*, 220-232.

Korkman, M., Kirk, U., Kemp, S. L. (1997). *NEPSY*. San Antonio, TX: The Psychological Corporation. (See also, Kemp, S. L., U. Kirk, & Korkman, M. *Essentials of NEPSY® Assessment*. John Wiley and Sons.)

LaFreniere, P.J., & Dumans, J. E. (1995). *The Social Competence and Behavior Evaluation—Preschool Edition*. Los Angeles, CA: Western Psychological Services.

Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (in preparation). *Preschool Comprehensive Test of Phonological & Print Processing*. Austin, TX: ProEd.

McCarthy, D. (1972). *McCarthy Scales of Children's Abilities*. San Antonio, TX: Psychological Corporation.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III: Technical manual*. Itasca, IL: The Riverside Publishing Company.

Molfese, D. L., Burger-Judisch, L. M., Gill, L. A., Golinkoff, R. M. & Hirsh-Pasek, K. A. (1996). Electrophysiological correlates of noun-verb processing in adults. *Brain and Language, 54*, 388-413.

National Center for Learning Disabilities (2002). *Get Ready to Read!: Screening Tool.* Lebanon, IN: Pearson Early Learning. (See also http://www.getreadytoread.org/research.html)

Pianta, R.C. (2001). *Student-Teacher Relationship Scale*. Luts, FL: Psychological Assessment Resources, Inc. (See also, Pianta, R.C. [2001]. *Student-Teacher Relationship Scale: Professional Manual.* Luts, FL: Psychological Assessment Resources, Inc.)

Reid, D., Hresko, W., & Hammill, D. (2001). *Test of Early Reading Ability Third Edition*. Austin, TX: ProEd.

Reynell, J. K. & Gruber, C.P. (1990). *Reynell Developmental Language Scales.* Los Angeles, CA: Western Psychological Services.

Reynolds, C.R., & Kamhaus, R. W. (1998). *BASC Behavioral Assessment System for Children: Manual.* Circle Pines, MJN: American Guidance Service, Inc.

Roid, G. H. and Miller, L. J. (1997). *Examiners manual: Leiter International Performance Scale-Revised*. Chicago, IL: Stoelting Co.

Shapior, L. R., & Hudson, J. A. (1991). Tell me a make-believe story: Coherence and cohesion in young children's picture-elicited narratives, *Developmental Psychology, 27*, 960-974.

Shure, M. B. (1992). *Preschool Interpersonal problem-solving (PIPS): Test manual.* (mshure@drexel.edu).

Shure, M. B. (1996). Interpersonal cognitive problem-solving: Primary prevention of early high risk behaviors in the preschool and primary years. In G. W. Albee & T. P. Gullotta (EDS.). *Primary Prevention Works* (pp. 167-188). Thousand Oaks, CA: Sage.

Sroufe, L. A., Schork, E., Motti, F., Lawroski, N., & LaFreniere, P. (1984). The role of affect in social competence. In C. E. Izard, J. Kagan, & R. B. Zajonc (Eds.), *Emotions, Cognition, & Behavior* (pp. 289-319). Cambridge: Cambridge University Press.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale; Fourth Edition. Guide for administering and scoring*. Itasca, IL: The Riverside Publishing Company. (See also Thorndike, R. L., Hagen, E. P., & Sattler, J. M. [1986]. *Stanford-Binet Intelligence Scale; Fourth Edition Technical manual.* Itasca, IL: The Riverside Publishing Company.)

Wagner, R. K. and Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*, *101*, 192-212.

Wagner, R., Torgesen, J., & Rashotte, C. (1999). *Comprehensive Test of Phonological Processing*. Austin, TX: ProEd.

Waters, E., Vaughn, B., Posada, G. & Kondo-Ikemura, K. (Eds.). (1995). Caregiving, cultural and cognitive perspectives on secure-base behavior and working models: New growing points of attachment theory and research. *SRCD Monographs*, vol. 60, nos. 2-3, (Serial no. 244). Pp. 280-282. The Attachment Q-set may be reviewed at http://www.psychology.sunysb.edu/ewaters/measures/aqs.htm; http://www.psychology.sunysb.edu/ewaters/measures/aqstext.htm.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intellginece—Revised (WPPSI-R): Manual.* San Antonio, TX: The Psychological Corporation.

Werthamer-Larsson, L., Kellam, S.G., & Wheeler, L. (1991). Effect of first-grade classroom environment on child shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology, 19*, 585-602.

Whetherby, A.M. & Prizant, B.M. (1993). *Communication and Symbolic Behavior Scales.* Baltimore, MD: Brooks Publishing Company.

Wiig, E., Secord, W., & Semel, E. (1992). *Clinical Evaluation of Language Fundamentals-Preschool*. New York: The Psychological Corporation.

Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests – Revised.* Circle Pines, MN: American Guidance Service (AGS) Publishing.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III*. Itasca, IL: Riverside Publishing.

Zimmerman, I.L., Steiner, V. G., Pond, R. E. (2002). *Preschool Language Scale-Fourth Edition.* San Antonio, TX: The Psychological Corporation.

Appendix A
Participants

Invited Participants:

Martha Abbott-Shim, PhD
Georgia State University

Marilyn J. Adams
Soliloquy Learning

Clancy Blair, PhD
Pennsylvania State University

Sarah J. Brainard
Stamford Child Care Center

Margaret (Peg) Burchinal, PhD
University of North Carolina – Chapel Hill
FPG Child Development Institute

Douglas Clements, PhD
State University of New York, University of Buffalo

Richard (Dick) M. Clifford, PhD
University of North Carolina at Chapel Hill
National Center for Early Development & Learning

Ronna Cook
Westat, Inc.

Anne E. Cunningham, PhD
University of California, Berkeley

Gayle Cunningham, MS
Jefferson County Committee for Economic Opportunity

Susanne Denham
George Mason University

David Dickinson, PhD
Education Development Center

Herbert P. Ginsburg, PhD
Columbia University, Teachers College

Susan L. Golbeck, PhD
Rutgers University

Mark Greenberg, PhD
Pennsylvania State University

Carol S. Hammer
Pennsylvania State University

Kathryn A. Hirsh-Pasek
Temple University

Marilou Hyson, PhD
National Association for the Education of Young Children

Nicholas S. Ialongo, PhD
Johns Hopkins Bloomberg School of Public Health

Aquiles Iglesias, PhD
Temple University

Sharon L. Kagan, EdD
Columbia University, Teachers College

Ann P. Kaiser, PhD
Vanderbilt University

Susan H. Landry
University of Texas – Houston

Jean Layzer
Abt Associates, Inc.

Mark W. Lipsey, PhD
Vanderbilt Institute for Public Policy Studies

Christopher J. Lonigan, PhD
Florida State University

John M. Love, PhD
Mathematica Policy Research

Scott R. McConnell, PhD
University of Minnesota

Ruth Hubbell McKey, PhD
Xtria, Inc.

Victoria Molfese
University of Louisville

Robin Morris, PhD
Georgia State University

Frederick J. Morrison, PhD
University of Michigan

Laura-Ann Petitto, PhD
Dartmouth College

Eva M. Pomerantz, PhD
University of Illinois, Urbana-Champaign

Kenneth Pugh, PhD
Yale University School of Medicine &
Haskins Laboratories

Craig T. Ramey, PhD
Georgetown University

Sharon Ramey
Georgetown University

C. Cybele Raver, PhD
University of Chicago

JoAnn L. Robinson, PhD
University of Colorado

Hollis S. Scarborough, PhD
Haskins Laboratories

Prentice Starkey, PhD
University of California, Berkeley

Dorothy S. Strickland
Rutgers, The State University of New Jersey

Doug Tynan, PhD, ABPP
AI duPont Hospital for Children

Cheri Vogel, PhD
Mathematica Policy Research

Catherine Walsh, MPH
Rhode Island KIDS Count

Jerry West, PhD
National Center for Education Statistics

Martha Zaslow, PhD
Child Trends

Nicholas Zill, PhD
Westat, Inc.

U.S. Department of Health and Human Services:

NICHD:
G. Reid Lyon, PhD
Child Development & Behavior Branch

Peggy McCardle, PhD, MPH
Child Development and Behavior Branch

Kyle Snow, PhD
Child Development and Behavior Branch

Melissa Welch-Ross, PhD
Child Development and Behavior Branch

Tanya Shuy
Child Development and Behavior Branch

Vinita Chhabra, Med
Child Development and Behavior Branch

Marita Hopmann, PhD
Division of Scientific Review

Natasha Cabrera, PhD
Expert in Child Development

ACF:
Wade F. Horn, PhD
Assistant Secretary, Administration for Children and Families

Joan Ohl
Commissioner, Administration on Children, Youth and Families

Windy Hill
Associate Commissioner, Head Start Bureau

Shannon Christian
Commissioner, Child Care Bureau

Naomi Goldstein
Division of Child and Family Development
Office of Planning, Research and Evaluation

Michael L. Lopez, PhD
Child Outcomes Research & Evaluation Office
of Planning, Research and Evaluation

Louisa Tarullo, EdD
Child Outcomes Research & Evaluation
Office of Planning, Research and Evaluation

Rachel C. Cohen, PhD
Child Outcomes Research & Evaluation
Office of Planning, Research and Evaluation

Mary Bruce Webb, PhD
Child Outcomes Research & Evaluation
Office of Planning, Research and Evaluation ED:

Helen Raikes
SRCD Visiting Scholar

Jim O'Brien
Head Start Bureau

Tom Schultz
Program Support Division, Head Start Bureau

Ivelisse Martinez-Beck, PhD
Policy Fellow, Child Care Bureau

NIMH:
Cheryl A. Boyce, PhD
Developmental Psychopathology and
Prevention Research Branch

ASPE:
Martha Moorehouse, PhD
Division of Child & Youth Policy

Denise Bradley, PhD
Division of Child & Youth Policy

U.S. Department of Education:

Gail R. Houle, PhD
Office of Special Education Programs

Heidi Schweingruber, PhD
Office of Educational Research and
Improvement

Appendix B

Language Milestones and Associated Constructs and Measures

	"Pre" Language ***Birth - 12 months***	**Language Onset** *12 months - 24 months (2 years)*	**Language Growth** *2 years – 3 years*	**Language Growth & onset of Meta-Language Awareness** Pre-Reading *3 years – 4 years*	**Reading** *4 years – 5 years*
Universal Language Milestones/Capacity at this age (Outcome) and Literacy (below)	Language Milestones Lang Perception • discriminates phonemes in (and segments) speech stream Lang Production • babbling Motor • physical/social growth, including reaching, grasping & showing; walking Social • joint attention,social pragmatics (e.g., communicative gestures w/multiple intents) • rudimentary conversational structure (vocalizes when adult is silent, silent when adult vocalizes)	Language Milestones **First word** milestone around 12 mths; **First 2-word** milestone around 18 mths, **First 50 words** around 18-24 mths (also true for early bilinguals in each of their languages); "**Vocabulary spurt"** around 18-22 mths (including growth in semantic richness of words)	*Language Milestones* **Vocabulary growth** • plus semantic growth Phonological •production of most of sounds of the language **Morphological** • marking on basic words appear • onset of over-regularizations Syntactic • length of basic sentences increase • growth in phrasal, clausal structure complexity **Story Telling** • very basic capacity appears **Conversation** • structure/pragmatic complexity increases over time	*Language Milestones* **Vocabulary growth** • plus semantic growth **Phonological, Morphological, and Syntactic growth** • comprehension & production of more complex syntax (including relative clauses, and passive sentences) **Rhyming/Word-Play** • word-play/humor (jokes, puns) with language appears **Stories=>Narratives** **Conversational**…	*Language Milestones* **Vocabulary growth** • Later developing phonology refinements and complex syllables, plus semantic growth Morpho & Syntactic • basic morpho & syntactic knowledge now stabilized in native language, with additional "later-syntax" embellishments Narratives flourish ***Conversational…***

	"Pre" Language *Birth - 12 months*	**Language Onset** *12 months - 24 months (2 years)*	**Language Growth** *2 years – 3 years*	**Language Growth & onset of Meta-Language Awareness** Pre-Reading *3 years – 4 years*	**Reading** *4 years – 5 years*
Assessments	**Language** • Universal Hearing Screening • Phonemic Discrimination Task **Social** • Social/ Conversational: Turn-Taking & Joint Attention **Motor** • Motor Milestones **Neuroscreening** -ERP left-hemisphere (LH) laterality test for speech processing -Mismatch Odd-Ball task (Pugh) -LH Laterality of Vocal Babbling (Holowka &Petitto)	• **MacArthur CDI Non-Verbal** • MacArthur CDI Vocabulary (Comprehension + Expressive)	• Reynell Plus (vocabulary and grammar) -receptive -expressive •Stories/Narratives: "Frog Story" (Produce & Repeat) • Conversational (Instrumental Language) * Phonetic Inventory (error types, percent Consonants/Vowels correct) *Biemiller-vocab *Word Order	• Reynell Plus (vocabulary and grammar) -receptive -expressive • Stories/Narratives: Shortened "Frog Story" (Produce & Repeat) • Conversational… •Woodcock-Johnson Letter-Word Identification *Modified Story and Print Concepts *Rhyme and Deletion tasks of the Early Phonemic Awareness Profile *Productive Phonology (Bankson-Bernthal Test of Phonology) *Biemiller-vocab * Dynamic indicators of basic early literacy skills (Dibels)"	• Reynell Plus (vocabulary and grammar) -receptive -expressive • Stories/Narratives: Shortened "Frog Story" (Produce-Repeat) • Conversational… • **Woodcock-Johnson Letter-Word Identification** *Woodcock-Johnson Letter-Incomplete Word • Letter identification (with mixed letters) • Writes letters, writes name, reads name •Observation of parent/child book reading w/CLDES *Biemiller-vocab *Quality of phonological knowledge (Complex Productive Phonology, & Assessment of Phonological Processes-R-complex word list) *Dibels
Who Administers?	Specialists	Parents	Teacher; *=Speech-language Therapist	Teacher; *=Speech-language Therapist	Teacher; *=Speech-language Therapist

**Note*. Kathy Hirsh-Pasek and Laura-Ann Petitto prepared this working table for workshop discussion.