B. SPECIFIC AIMS

Research has demonstrated that much of the academic achievement gap observed in economically disadvantaged children is already present when formal schooling begins at age 5. This fact has focused the attention of policy-makers and researchers on educational and enrichment programs for children from birth to age 5, particularly on Head Start and pre-kindergarten programs offered to children and families with economic and other disadvantages. The goal of these programs is to bring disadvantaged children up to full school readiness when they enter kindergarten. Evidence suggests that these programs improve school readiness skills, but that they have heterogeneous effects. This raises questions about which subsets of children experience larger and smaller program impacts. Determining the differences in intervention efficacy for children from different backgrounds and/or with different characteristics is necessary to meet the goal of creating programs ensuring that most children enter kindergarten ready to learn.

As with the other subprojects in this program project, we generate hypotheses by invoking an important framework from developmental science – that of person-environment (for us, child-policy) and stage-environment (stage-policy) fit. The period from birth to five involves rapid developmental change, often combined with stressful environmental change. As children develop oral language, reasoning, and early numeracy, as well as attention and behavioral skills during this period, most transition from time spent primarily with the family to substantial periods of time in child care and larger peer and instructional groups outside the home. The same issues of child-policy and stage-policy fit that have proven useful in understanding adolescent development are also useful in understanding developmental outcomes from birth to age five.

Using experimental and regression discontinuity intervention studies as "environments," we will employ the child-policy and stage-policy fit perspectives to derive and test hypotheses about which combinations of child, family, and child care program characteristics lead to better and worse cognitive/achievement and emotional-behavioral readiness outcomes. We will test a series of hypotheses using random-assignment data from the Infant Health and Development Project (IHDP), the Early Head Start Study (EHS), the Head Start Impact Study (HSIS), and the randomized Preschool Curriculum Evaluation Research Initiative (PCER). We will also analyze data from the regression-discontinuity study of the Oklahoma Universal Pre-K program. When data from the HSIS are used to test these hypotheses, we will reconcile our subgroup-based results with those from Project III's use of distributional methods. In addition, we will estimate the effects of particular program characteristics (e.g., standard global measures of program quality -- ECERS/FDCRS; a standard measure of the quality and nature of teacher sensitivity to students – the Arnett; TBRS measures of the extent and nature of instruction in literacy and math; whether the center uses a curriculum; the child/staff ratio, center size, teacher turnover; and teacher demographic, education and training characteristics) for particular subgroups of children. We will use an instrumental variables strategy that employs the variation in program characteristics due to the experimental treatment to provide unbiased estimates of the effects of these characteristics on child outcomes. Specific research aims:

Aim 1. To test the *compensatory hypothesis* that, owing to their greater environmental vulnerability, high-risk children benefit the most from high-quality birth to five programs. "Higher risk" is operationalized as very low income and/or experiencing many family risk factors or low cognitive performance at entry to the program.

Aim 2. To test the *skill begets skill hypothesis*. This predicts that children who have better skills at entry will be able to learn more quickly and gain more from high-quality programs, which is opposite to the predictions regarding child skill/program quality interactions from the compensatory hypothesis.

Aim 3. To test the *protective* and *cumulative disadvantage hypotheses* that poor-quality programs are less detrimental for children whose personal or family characteristics protect them from poor environments, and are most detrimental for children with a high number of risk factors.

Aim 4. To test the *differential susceptibility* hypothesis that, for children with difficult temperaments, high-quality child care magnifies positive impacts *and* low-quality care magnifies negative impacts.

Aim 5. To determine which particular preschool program characteristics best boost school readiness effects for important population subgroups by using instrumental variables estimation on random assignment datasets (HSIS and PCER) containing across-classroom variation in these program institutional, instructional and other characteristics.

C. RESEARCH STRATEGY

C.1 Significance

Identifying which early childhood programs are effective for which children is a pressing policy issue. Fully three-fourths of children under 5 years are in child care on a regular basis (U.S. Census Bureau, 2006). Publicly-funded parent education and home-based interventions serve an estimated 1.1 million children in prekindergarten programs (Barnett et al., 2008) and 900,000 children in Head Start programs (Office of Head Start 2008). Research and evaluation of these center-based and home-based programs have largely focused on main effects, asking the question "what is the overall effect of these programs?" In the proposed study, we ask a more nuanced, hypothesis-driven set of questions regarding which programs serving particular groups of children yield larger effects on particular sets of cognitive-academic and social-behavioral outcomes.

Several theoretical models guide this study. A central proposition of bioecological theory (Bronfenbrenner & Morris, 2006) and life course theory (Elder, 1996) is that experiences have differential impact depending on the individual's developmental status and personal and family characteristics. Both human and animal studies highlight the critical importance of early childhood for brain development (Knudsen et al., 2006), while some economic models of human capital development (e.g., Cunha, Heckman, Lochner & Masterov, 2006) presume that preschool cognitive and social-emotional capacities are key ingredients for success during the school years.

Social cultural development theories of learning (Vygotsky, 1978) focus on the quality of the match between a child's background and skill level and the level and quality of instruction, hypothesizing that learning occurs when quality instruction is only slightly above the child's skill level. These theories combined with infant behavioral and brain development research, have led to a science of early child development that includes the following four principles (Heckman, 2006; Shonkoff & Phillips, 2000): (1) development is influenced by the interaction between genetics and environmental experiences; (2) skill development is hierarchical, with later abilities building on the mastery of earlier skills; (3) developmental competencies in cognitive, language, social, and emotional domains are interdependent and central to success in school and beyond; and (4) although adaptation is a lifelong process, brain circuitry and associated behaviors develop primarily during sensitive early periods of life. The ramifications of these principles are twofold: first, to help ameliorate the negative effects of multiple risks associated with poverty on children's development, interventions must begin early in children's lives and second, these interventions must match children's developmental needs and be sufficiently intensive to achieve their goals (Ramey & Ramey, 1999). Thus, program impacts will vary depending on both the child's background and the extent to which instruction is high quality and well-matched to the child's skill level.

C.1.a Hypotheses

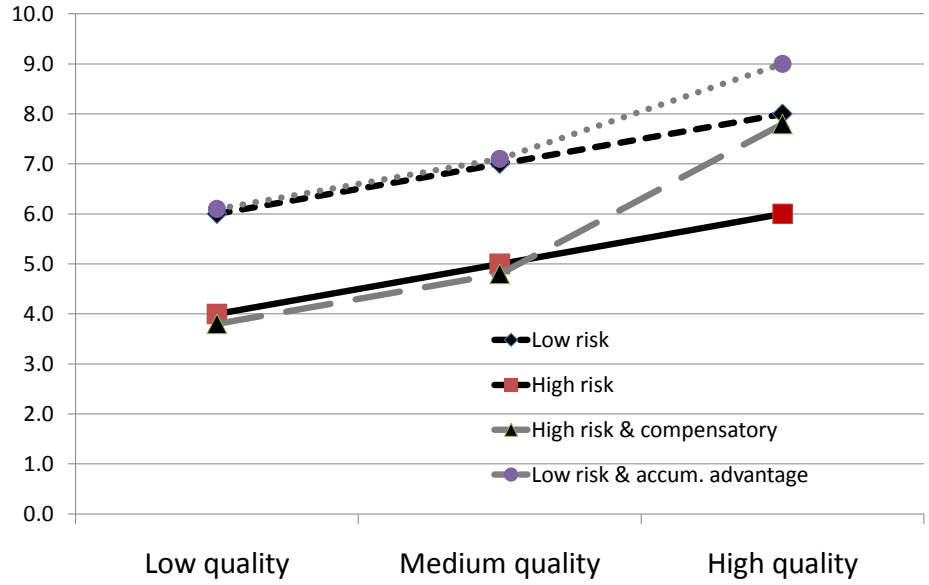
Our overall hypothesis is that high-quality early childhood programs do not affect all children to the same degree. Thus, the impacts on children's outcomes differ as a function of family resources, child cognitive performance at baseline, age and temperament. This departure from a one-size-fits-all formulation of program effects is expected to reveal the conditions under which early childhood policies are more (or less) effective. We treat impact heterogeneity as a feature of programs that needs to be conceptualized and tested.

Prior researchers have postulated several competing hypotheses about differential program effects. Two of these are relevant to children's participation in high-quality early education programs and specify who is expected to derive greater benefit from these high-quality programs. The *compensatory* hypothesis (Sameroff & Chandler, 1975) predicts that children who are at risk because of economic disadvantage, low skills, or difficult temperaments derive greater benefit from high-quality early education programs relative to children who are not at risk. This hypothesis provided the rationale for the funding of programs such as Head Start. Alternatively, *accumulated advantages* and *skill begets skill* (Cunha et al. 2006) posit that children with greater initial individual abilities (*skill begets skill*) or less-risky advantage-laden family environments (*accumulated advantages*) will derive greater benefits from high-quality early education programs than less advantaged peers because of their ability to build on existing skills or family advantage.

Figure C.1 depicts the *compensatory* hypothesis. Child care quality runs along the X axis while some valued child achievement or social-emotional outcome runs along the Y axis. The parallel lines for high- (solid) and low- (short dashed) risk children reflect assumed modest and parallel boosts in child outcomes as child care quality improves. The compensatory hypothesis (shown with the non-linear long dashed line) presumes that the fit of a well-designed high-quality program to the needs of high risk children produces positive benefits that may reduce or even be sufficient to eliminate the outcomes gap between higher and lower-risk children.

In contrast, the *accumulated advantages* hypothesis presumes that the higher skills of lower-risk children will increase the productivity of investments like child care quality and impart a steeper slope for low-risk relative to high-risk children. In terms of Figure C.1, a steeper slope to the low-risk relative to high-risk line would produce a larger between-group gap in the high-quality condition. A special case of *accumulated advantages* hypothesis is the *skill begets skill* hypothesis, in which attention focuses on the child's own baseline cognitive and noncognitive skills. Cunha and Heckman (2008) develop a model that includes the cumulative role of cognitive and noncognitive skills, as well as skill investments made by families, preschool programs and schools. Suppose we distinguish the periods birth to age 3 (period 1), and ages 4 to 6 (period 2). At birth (period 0), children have endowments of cognitive potential and temperament (S^C_0, S^N_0) that reflect some combination of genetic and prenatal influences. School readiness human capital (h) is a product of an individual's eventual period 2 cognitive and noncognitive skills: $h = g(S^C_2, S^N_2)$. Achievement-related skills in period t are a product of cognitive and noncognitive skills in the prior period, plus current investments. We focus on how skills acquired during $t = 2$ relate to both the achievement and noncognitive skills (S^C_1, S^N_1,) children bring to preschool, plus the interaction between those start-of-preschool skills and the preschool investments themselves. This interaction, in which higher skill at kindergarten entry enhance the productivity of preschool investment identifies the "skill begets skill" effect.

Figure C.1: Compensatory hypothesis for high risk kids



Other hypotheses focus on differential effects of *poor-quality* programs on child outcomes. This *cumulative risk* hypothesis (Sameroff & Chandler, 1975) predicts that poor-quality programs are particularly detrimental for children already at risk because of economic disadvantage, low birth weight, difficult temperament, or other reason. Alternatively, the *protective hypothesis* (Garmezy, 1985) predicts that poor-quality programs are less detrimental for children whose personal qualities or family resources serve to protect or offset their negative impacts. Using experimental databases, we will test each of these hypotheses regarding child by program quality interactions. We will also test an intriguing fifth hypothesis, *differential susceptibility*. This hypothesis, first documented in animals (Suomi, 1995) and recently proposed for humans by Belsky and colleagues (Belsky, 1997, 2005; Belsky et al., 2007), posits that some individuals respond more sensitively to negative *and* to supportive environments. Belsky has observed that temperamentally difficult children are more negatively affected by poor-quality environments AND more positively affected by high-quality environments.

C.1.b Family Economic and Educational Resources

An important policy question is whether high-quality early childhood programs are particularly beneficial and poor-quality programs are particularly problematic for children whose families have few economic and educational resources. Although some poor children attend federally funded Head Start programs, state-funded pre-K programs, and various locally supported programs, the number of available slots is less than the number of eligible children (Barnett et al., 2008). Access to high-quality programs is particularly limited for children whose families' incomes are just over the income requirements (Dowsett, Huston, Imes, & Gennetian, 2008). We ask: Are there differential effects of high- and low-quality programs for poor and near-poor children versus middle class children? There is non-experimental evidence that poor children are more susceptible than middle class children to variations in child care quality, with sensitive and stimulating child care buffering children from the effects of poverty, perhaps by compensating for poor-quality home environments. For

example, McCartney and colleagues (2007) found significant interactions between high-quality care and family income. Low-income children receiving high-quality care displayed greater school readiness than those who received lower quality or no formal care. Children in low-quality care performed better than those with no formal care with respect to receptive and expressive language. Nonexperimental research by Burchinal and colleagues (2006) also found that higher quality child care moderated the negative association between family risk, including income, and children's academic skills in elementary school.

The IHDP and Oklahoma data are particularly well suited to the examination of family resources x program interactions because they include economically diverse households of advantaged as well as poor and near-poor families. We will examine family resources x program interactions in the EHS and HSIS samples by contrasting effects on children of very disadvantaged families with children of somewhat less disadvantaged (but still poor) families. There is some evidence that family resources x program interactions may occur within this more truncated range (Love et al., 2005). Thus:

Hypothesis 1*:* Consistent with the *compensatory hypothesis,* the effects of high-quality child care programs on cognitive-achievement and emotional-behavioral outcomes are expected to be greater for children from "high risk" family backgrounds (very low SES and many family risk factors) than for children from more advantaged family backgrounds. This is contrasted with the *accumulated advantage hypothesis* in which high SES and the absence of other family risk factors provide the scaffolding that enables children to take the best advantage of high-quality programs. The compensatory hypothesis's prediction regarding bigger program payoffs for the lowest-skill children stands in contrast to the *skill begets skill hypothesis*, which predicts that children with the highest individual skills will profit the most from high-quality child care.

Hypothesis 2: Consistent with the *protective hypothesis*, the effects of poor-quality programs are less detrimental for children whose personal or family characteristics protect them from poor-quality environments. (This differs from H1 in focusing on the effects of *poor-quality* rather than *high-quality care.)*

C.1.c Child Age or Developmental Timing

A second policy question is whether gains from early childhood programs vary as a function of the children's age. This is likely because infants, toddlers, and preschoolers have different capacities and needs. During infancy, children are forming attachment relationships with parents and caregivers. Caregiving demands are high and often require one-on-one attention (Steinberg, Vandell, & Bornstein, 2009). For toddlers, language development and emotional regulation are key issues (Steinberg et al., 2009). For preschoolers, cooperative play with peers, pretend play, and emerging numeracy and literacy skills are important (Steinberg et al, 2009). These age-related demands have implications for age-appropriate programming. Further, intervening earlier should provide the social, cognitive, and attention skills that are important for later development. Opinions vary regarding the age at which children should begin programs with a formal educational agenda, but this debate often ignores the issue of center quality. Some researchers (Loeb & Lee, 1995) believe that these experiences are appropriate for 4- and 5-year-olds, but not younger children. Others (Helburn et al., 1994) argue that center care is beneficial for 2- and 3-year-olds, but not infants.

In the proposed study, three of the datasets can be used to test hypotheses that early childhood program effects are moderated by child age in combination with family resources. The Head Start Impact Study (HSIS) randomly selected programs, and studied 3- and 4-year-old children who were randomly assigned to attend Head Start or not. This will provide the best test of the hypothesis. The other studies provide indirect evidence by drawing from target populations with baseline ages that differed across studies. The Oklahoma Pre-K Study will enable us to compare 5-year-olds who had completed one year of pre-K with peers without pre-K experience. The Early Head Start Study (EHS) consists of either a center-based intervention or combination of home visits and center-based care. One-third of the children received the center-based early childhood program beginning in infancy and continuing to age three, whereas two-thirds of the families received home-visiting in the first year and center type care in the second and third years.

Hypothesis 3: Consistent with the *compensatory hypothesis,* because younger preschoolers are likely to be more environmentally vulnerable than older preschoolers, we expect larger effects of high-quality programs for younger children. For children who are in the very high risk group, the stronger treatment of high-quality center type care in infancy is expected to yield better outcomes than home visiting alone.

C.1.d Child Temperament

Research suggests that child emotional negativity and effortful control moderate the effects of early

childhood programs on child outcomes. Pluess & Belsky (2009) found that effects of early child care quality on behavior problems and social competence were moderated by infant negativity. Consistent with the differential susceptibility hypothesis, children with difficult temperaments were more affected by child care quality than were temperamentally easy children. Children who were temperamentally difficult as infants exhibited more subsequent behavior problems when they had low-quality child care *but* exhibited fewer behavior problems when they had high-quality care. Children who were temperamentally easy as infants appeared less susceptible to child care quality. Four of the proposed datasets (IHDP, EHS, HSIS, and PCER) include measures of child emotionality that can be used to test the differential susceptibility hypothesis.

Hypothesis 4: Consistent with the *compensatory hypothesis,* the effects of high-quality child care programs on cognitive-achievement and emotional-behavioral readiness are expected to be greater for children with difficult temperaments.

Hypothesis 5*:* The *differential susceptibility hypothesis* extends the compensatory hypothesis that high-quality child care has a particularly strong positive effect on children with difficult temperaments to add the hypothesis that for such children, low-quality child care has particularly strong negative impacts on cognitive-achievement and emotional-behavioral readiness.

C.1.e Seeing Inside the Black Box

Our five hypotheses require distinguishing high versus lower quality child care. In addition, we will analyze the HSIS and PCER experiments, in which children were randomly assigned to treatment programs which induced variation in center environment, teacher/care provider qualifications and training, classroom environment, and classroom activities. Looking across centers and classrooms, using multilevel and instrumental variable modeling techniques (Riccio & Bloom 2002; Morris, Duncan, & Clark-Kaufman 2005; Kling, Liebman, & Katz 2007; Bloom and Zhu 2010), we will test whether specific program characteristics (e.g., standard global measures of program quality -- the ECERS/FDCRS; a standard measure of the quality and nature of teacher sensitivity to students – the Arnett; TBRS measures of the extent and nature of instruction in literacy and math; whether the center uses a curriculum; the child/staff ratio, center size, teacher turnover; and teacher demographic, education and training characteristics) are particularly successful with students with varying characteristics (e.g., the most or least at-risk).That is, we will use an instrumental variables strategy that employs the variation in program characteristics due to the experimental treatment to provide unbiased estimates of the effects of these characteristics on child outcomes.

C.2 Innovation

Our project is innovative along at least five dimensions. (1) It is the first to comprehensively test a complete set of hypotheses regarding the effect of the fit between child care quality and child at-risk status on child outcomes. (2) It uses five different datasets, which increases robustness compared to prior studies. (3) Five of these involve random-assignment experimental data; the other (Oklahoma) uses a quasi-experimental regression-discontinuity design. These designs reduce possible bias due to omitted variables (such as genetics). (4) These databases include experiments (HSIS, PCER) inducing variation in measured program features to "get inside the black box" and discover which program elements are most successful with which particular at-risk students. (5) The analyses will employ up-to-date statistical techniques, including multilevel modeling, instrumental variables combined with randomized data, and statistical meta-analysis.

C.3 Approach

C.3.a Data

The *Infant Health and Development Program* (IHDP) provided services to infants and toddlers. This was an eight-site randomized clinical trial designed to test the efficacy of a comprehensive early intervention program for low birthweight (LBW) premature infants. Infants weighing 2,500 g (5.51 lb) or less at birth were screened for eligibility if their postconceptional age between January and October 1985 was 37 weeks or less and if they were born in one of eight participating medical institutions. A total of 985 infants were randomly assigned either to a medical follow-up only or to a comprehensive early childhood intervention. The treatment group received weekly home visits through 12 months of age, consisting of parenting education, mental health counseling and support, and referral to social services. Further, from ages 1 to 3, treatment group children were eligible to attend a free, full-day, high-quality child development center run by IHDP staff. Assessment ages for the IHDP are 1, 2, 3, 5, 8 and 18 years. Program cognitive and achievement effect sizes were .75 at 3 years of age but had declined to zero by age 18. The evaluation of the *Early Head Start* (EHS) program provides a second database covering an experimental, quality child care intervention for infants/toddlers. In 1996, 3,001 children

under 1 year of age were randomly assigned to receive EHS services, or to a control group. Treatment (main) effect sizes were .13 on language and .10 on adjustment at 3 years for the entire program (Love et al., 2005). Data were collected when the children were 14, 24, 36, and 60 months old. Data included direct assessments of the children, laboratory tasks, maternal interviews, and observations of the home and child care environments. These data can reveal the consequences of random assignment to good-quality infant/toddler child care services, combined with nonexperimental variation in quality, quantity, and type of preschool care.

Neither the IHDP nor EHS treatments were provided beyond the child's third birthday; however both studies collected follow-up data during the preschool period. For treatments administered during preschool ages, we will use data from the *Head Start Impact Study* (HSIS) and the *Oklahoma Universal Pre-K Study*. In the HSIS, a total of 4,667 three- and four-year-olds were randomly assigned from waiting lists for a clustered, nationally-representative sample of Head Start centers to receive Head Start services or to a control group. Assessments included direct assessments of achievement and language skills and parent and teacher interviews. In addition, the HSIS database includes a wide array of detailed program characteristics. These include standard child care quality measures (ECERS for centers and FDCRS for home-based care), the Assessment Profile for Classrooms, measures of classroom instruction, and the Caregiver Involvement Scale (CIS). Information is also provided on adult-child ratios and class size. Child care providers describe their education and years of experience, program elements such as curriculum and program support, quality of management, and literacy and math activities. Center directors provided information on staffing, teacher education, etc. The *Oklahoma Universal Pre-K Study* focuses on preschoolers. This study used a quasi-experimental, regression-discontinuity design, relying on a strict birthday eligibility criterion to compare 1,461 "young" kindergarten children who just completed pre-K to 1,567 "old" pre-K children just beginning pre-K. The data report on the school readiness of children who attended the universal pre-K program in Tulsa, Oklahoma during the 2002-2003 school year. Family income varies widely, but measures of temperament or test scores prior to program entry are absent, so these data cannot be used to test the *skill begets skill* or *differential susceptibility* hypotheses. However, they do permit testing the *compensatory* hypothesis.

The *Preschool Curriculum Evaluation Research Study* (PCER) focuses on preschoolers. It involved experimental evaluation of 14 preschool curricula in 12 sites from 2003 to 2005. In each of the 12 sites, preschool students or preschool programs were randomly assigned to either the treatment curriculum or a business-as-usual control condition. The study included data on 2,911 children in 315 preschool classrooms and 208 preschools. Five major data collection instruments included a child assessment, a teacher report, classroom observation, a teacher interview or questionnaire, and a parent interview (Preschool Evaluation Research Consortium 2008). The classroom observation provided a measure of program quality.

C.3.b Measurement of Variables in Each of the Data Bases

The primary measures are summarized in Table 1 (end of document). They are briefly described below.

C.3.b.1 Cognitive/achievement outcomes. The *Peabody Picture Vocabulary Test* (PPVT) was administered at entry to or during preschool in the IHDP, EHS, HSIS, and PCER studies. The PPVT is an achievement test of receptive vocabulary that relates to other measures of language, literacy, and academic achievement for children 2 years and older. The *Wechsler Preschool and Primary Scale of Intelligence* (WPPSI) was administered in the IHDP. It is a widely used measure of intelligence for children 4 to 6 years of age that exhibits good test-retest reliability, $r = .93$ to .96. The WPPSI Verbal IQ is based on performance on scales measuring skills such as verbal abstract reasoning and receptive and expressive vocabulary The *Bayley Scales of Infant Development* (IHDP, EHS) is a measure of general cognitive skills for infants under the age of 3 years. It assesses sensory perceptual acuity and discriminations; memory, learning, and problem solving; verbal communication; and the ability to form generalizations and classifications. Test-retest reliability for the mental and motor scales (.83 and .77, respectively) are high for scales of infant development. The Verbal Reasoning score on the *Stanford-Binet Intelligence Scale, Fourth Edition* (IHDP at 36 months) measures verbal knowledge and understanding obtained from the school (child care) and home learning environments and reflects the ability to apply verbal skills to new situations. In addition, for phonological awareness the PCER administered the Elision subtests of the *Preschool Comprehensive Test of Phonologic and Print Processing* and the *Comprehensive Test of Phonological Processing for Kindergarten* (Pre-CTOPP and CTOPP). PCER also administered the *Test of Language Development* (TOLD).

To measure academic achievement the EHS, HSIS, Oklahoma and PCER studies administered two subtests from the *Woodcock-Johnson Tests of Achievement-III* (WJ), the Letter-Word Identification subtest to measure early reading skills and the Applied Problems Subtest to measure early math skills. These studies used

the Woodcock-Johnson III, whose reliability coefficient for the 3- to 5-year-old age group ranges from .97 to .99 for Letter-Word and .92 to .94 for Applied Problems. In addition, PCER administered the *Grammatic Understanding* subtest, the *Child Math Assessment-Abbreviated* (CMA-A) and the *Building Blocks' Shape Composition Task*. The *Peabody Individual Achievement Test* (PIAT) was administered in the IHDP.

C.3.b.2 Behavioral outcomes: Learning-related, externalizing and internalizing problems. The IHDP, HSIS, EHS, and PCER administered a variety of measures of behavioral outcomes related to learning, externalizing problems, and internalizing problems. The *Preschool Learning Behaviors Scale* (PLBS; HSIS, PCER) is a teacher-report measure of preschool children's approaches to learning. Teachers are asked to rate how often a child exhibits particular behaviors for 29 items, which are used to create three scale scores, Competence Motivation, Attention-Persistence, and Attitude Toward Learning with reasonable internal consistencies (*r* = .87, .88, .78, respectively). The *Leiter-R Attention Sustained Scale* (HSIS, EHS) includes four trials. In each trial, children are shown a target figure and then required to scan an array of similar figures and mark the targets as quickly as possible. This subtest has good internal consistency for the 4- to 5-year-old version (Cronbach's ☐ = .83) and good test-retest reliability of *r* = .85. The *Child Behavior Checklist* (CBCL; parents and teachers, IHDP, EHS) is a widely used measure of internalizing and externalizing behavior problems, and includes an Attention Problems subscale. The 28-item *Behavior Problems Index* (HSIS) assesses problem behaviors with items drawn from several other child behavior scales (e.g., CBCL). Items are rated on a 3-point scale. Two-week test-retest reliability was .92.

C.3.b.3 Child care quality. Quality of care was measured with several standard instruments across the studies. The *Early Childhood Environment Rating Scale-Revised* (ECERS-R; EHS, HSIS, PCER) is a widely used measure of global classroom quality, specifically designed for use in classrooms serving children between 2½ and 5 years of age. Scores on the ECERS-R range from 1-7 with 1 indicating "inadequate" quality, 3 indicating "minimal" quality, 5 indicating "good" quality, and 7 indicating "excellent" quality. The scale's authors report a total scale internal consistency of .92. A comparable scale, the *Family Day Care Environment Rating Scale* (FDCRS; EHS, HSIS) was administered in child care homes. The *Caregiver Involvement Scale* (CIS; HSIS, PCER) is an observational scale consisting of 26 items reflecting teacher sensitivity, harshness, and detachment that are rated on a 1-4 scale indicating how characteristic they are of the teacher, from not at all (1) to very much (4). Psychometric analyses suggest a single factor most parsimoniously represented these data (Cronbach's ☐ = .93). In addition, PCER collected classroom observations and data on the fidelity of implementation of each curriculum.

C.3.b.4 Moderator (conditioning) variables. Parents reported their education levels, marital and employment status and family income and household size and composition in all of the studies. Maternal stressful events and depressive symptoms were measured in EHS, IHDP, and PCER. A variety of measures of the quality of the family environment are available on these databases. The quality of the home environment was measured in three studies (IHDP, EHS, HSIS) using the *Home Observation for Measurement of the Environment* (HOME). The HOME assesses overall quality of the physical and social resources available to the child within the home, through both direct observation and a semi-structured interview with the mother. Similar information was also collected from parents in the PCER. *Videotaped, semi-structured mother-child interactions* (e.g., Three/Two Bags Task) were collected in the EHS and rated for qualities of parenting (sensitivity to distress, intrusiveness, detachment, cognitive stimulation, positive regard for the child, negative regard for the child, and flatness of affect).

Child age at baseline averages 12 months in IHDP, spans 6 – 18 months in the EHS, spans 36 – 59 months in the HSIS, 48 – 60 months in the Oklahoma data, and 48 – 72 months in PCER. Child temperament at baseline is measured in the IHDP and EHS using the Bayley Infant Behavior Record. In the HSIS, parents report this information for their children at baseline (either 3 or 4 years of age). In PCER this was reported by parents. This information is not available in the Oklahoma database.

C.3.b.5 Control variables. Controls are described in Table 1 (end of document). Most reflect demographic characteristics and social risk factors collected during interviews with the mother. Mothers were also asked about their child's health and health history. Maternal depressive symptoms were measured with the *Center for Epidemiological Studies Depression Scale* (CES-D; HSIS, EHS, PCER), for which Cronbach's ☐ tends to be good, ranging from .88 to .95 in the SECCYD, and with the *General Health Questionnaire* (GHQ; IHDP), Cronbach's ☐ = .95.

C.3.b.6 Missing data, sample attrition. Information is missing on some variables for some cases. We will use multiple imputation for missing data, creating ten datasets for analysis, which will then be used to arrive at

a coefficient estimate and an appropriate standard error. Under the assumption that data are missing at random (MAR), which is likely to hold given the large number of controls available in these data bases, this methodology has been shown to produce estimates with excellent statistical properties (Allison, 2002). We have already used these methods successfully on our IES grant. In addition, we will use maximum likelihood estimation techniques that also have excellent properties when used with missing data.

C.3.c Methodological and Analytic Approaches

C.3.c.1 Aim 1 analysis. Aim 1 tests the *compensatory hypothesis* that, owing to their greater environmental vulnerability, high-risk children benefit the most from high-quality programs. "High risk" will be operationalized as low SES (compared to middle and higher SES) and many (as compared to few) family risk factors; and low (as opposed to higher) baseline cognitive performance. Evaluation evidence supports the proposition that the IHDP, EHS, HSIS, Oklahoma programs, and some of the PCER programs offered children a "high-quality" treatment relative to their control-group counterparts.

The resources and supportiveness of the family environment. Family SES provides a direct measure of family resources. Two of our datasets – IHDP and the Oklahoma Universal Pre-K program – offered services to children with a broad range of socioeconomic circumstances. Accordingly, they provide the strongest tests of the hypothesis of larger impacts for the lowest SES children, based on the model in which the treatment dummy T is interacted with key components of SES such as mother's education:

(1) Later Skills = $b_0 + b_1 T + b_2 SES + b_3 T \times SES + j$ (Controls) + u

The *compensatory hypothesis* posits a negative sign for coefficient b_3.

Because they restricted entry to poor or near-poor children, analogous tests of the compensatory hypothesis are limited in the case of the Early Head Start (EHS) evaluation data, and data from the Head Start Impact Study (HSIS). However, even within their low-income study samples, the EHS and HSIS have large enough sample sizes and sufficient variation in the components of SES (e.g., maternal education) that we will use them to test for interactions with treatment. Further, these databases include income/poverty threshold variation from below 1.0 to near 1.85, and we will also use this variation in our calculations.

In addition to SES, an index of cumulative social risk will be used. We will draw from the prior work of co-PI Burchinal, whose previous work discusses risk indices. Direct measures of family risk are available in all but the Oklahoma data. Risk scores will be computed from the baseline family interview using factors such as the mother's marital status, education, stressful events, psychological adjustment, and the family's income, household size, and employment status. This risk factor will be interacted with treatment. We hypothesize that treatments will be more effective for children with more risk factors as opposed to fewer risk factors.

Child's entry skills. Another element of baseline risk is low cognitive skill. Our tests here are confined to the HSIS and IHDP studies since Oklahoma Pre-K did not measure skills prior to program entry, and EHS began their treatments in the first year of life – a period in which cognitive skill cannot be measured reliably. (The IHDP also began at birth, but offered its high-quality center-based care beginning when the children were 12 months of age. In these data we can consider 12 months of age to be the baseline age without risking endogeneity bias, since there is no evidence of child impacts at 12 months [Brooks-Gunn et al., 2009]).

Our models will include treatment and baseline functioning main effects as well as interactions between medium child functioning-by-treatment dummies (T x MedIQ) and high functioning-by-treatment (T x HIIQ) dummies. The compensatory hypothesis predicts negative coefficients on these two interaction measures, with the largest discrepancy between the baseline low and high-skill groups.

(2) Later Skills = $b_0 + b_1 T + b_2 IQ + b_3 T \times MedIQ + b_4 T \times HIIQ + j$ (Controls) + u

As with our SES-based tests, we will control for treatment interactions with SES and other factors correlated with baseline cognitive skills in an attempt to isolate the skill-based interaction effects.

Child age. We hypothesize that younger preschoolers will benefit more than older preschoolers from high-quality interventions. By enrolling large samples of both 3- and 4-year-olds, the HSIS provides the best data to test this hypothesis. We will conduct this test by interacting the HSIS treatment dummy with the child's age at baseline. Impact results (Puma et al. 2010) support this hypothesis; our work will replicate these with formal tests of the interactions, and extend them to the longer-run outcome data. Narrow age ranges in the other studies preclude meaningful tests of treatment by child age interactions. A nonexperimental test of the child age hypothesis compares the treatment effect sizes across projects that provided similar interventions to children of different ages. The IHDP, HSIS, Oklahoma, and PCER programs all sought to provide high-quality child care. IHDP, Oklahoma, and PCER enrolled both low and high income families; HSIS and EHS enrolled

low-income families only. They varied in terms of the age at which the program began and the number of years the program was offered. IHDP began in infancy whereas the HSIS included 3 and 4 year-olds and Oklahoma Pre-K included only 4 year-olds. PCER included children from 4 to 6 years of age. Comparisons of treatment effect sizes, adjusting for years of treatment, show the consequences of intervening at different ages.

C.3.c.2 Aim 2 analyses. Aim 2 tests predictions from the *skill begets skill* hypothesis, which contradict those from the *compensatory* hypothesis. The latter predicts vulnerable groups profit more from high-quality child care than other groups; the former predicts that children from advantaged groups have the higher skills that enable them to profit most from high-quality child care. To tests these competing predictions we will rely most on the two studies with large sample sizes and baseline child skill assessments –IHDP and HSIS. Since the essence of skill begets skill is that high child-based *skills* at baseline enhance the productivity of high-quality care investments, skills, our interaction tests are essentially the same as in equation (2). The skills begets skill hypothesis predicts stronger treatment effects as one moves up the child-skill scale. Our interaction tests will be broader than those specified above by including an assortment of baseline cognitive skills (measured by the Bayley and PPVT), as well as baseline learning-related behaviors and externalizing and internalizing behavior problems (see Table 1 at end of document for measurement details).

A child-fit perspective broadens the simple skill begets skill model to allow for heterogeneous investments (that is, programs with differing curricular goals, techniques, and intensity levels), some of which (e.g., IHDP, Head Start, some pre-K programs) might be directed at boosting the basic skills of low-skill children and others (arguably K-12 education) directed primarily at normative development. Given the nature of the investment programs we study, we expect to find more support for the compensatory than skill begets skill hypotheses.

C.3.c.3 Aim 3 analyses. Aims 1 and 2 test hypotheses regarding the impacts of *high-quality* care. Shifting to focus on *low-quality* programs, the *protective hypothesis* holds that poor-quality programs are less detrimental for children whose personal or family characteristics protect them, while the *cumulative disadvantage hypothesis* posits that poor-quality programs will be particularly detrimental for high-risk children. Testing these hypotheses requires us to identify groups of children who experienced low-quality care. In the case of EHS, two kinds of quality splits are possible. Careful attention to measuring quality of implementation produced a dichotomy that distinguishes sites with higher vs. lower quality implementation. This split was used successfully in the Love et al. (2002) report to distinguish between sites with larger and smaller impacts on some of the short-run child outcomes. Second, the EHS programs can be distinguished by type – parent visitation only vs. a combination of parent visitation and center-based care. We take the former to constitute a medium quality treatment and the latter to constitute high-quality care.

To identify a low-quality care condition in the HSIS we rely on the fact that child care quality appears to be heavily clustered by geographic location (see below) and that HSIS has direct measures of Head Start program quality. Geographic clustering is likely to produce medium quality care environments, on average, for the control-group counterparts to children attending high-quality Head Start programs and, most importantly, low-quality care environments, on average, for the control-group counterparts to children attending lower-quality Head Start programs. We then consider the control groups for the below-median quality HSIS centers to be experiencing "predicted low-quality" care, while the treatment groups with below-median center-based quality scores constitute the counterfactual "predicted medium quality" care condition. This contrast provides a basis for testing our hypotheses involving low-quality care.

To implement the HSIS-based quality split, we rely on the fact that HSIS study observers rated all centers on the ECERS quality scale and all home-based child care on the FDCRS quality scale. We consider below-median ECERS score to indicate Head Start centers providing "medium quality" treatments, with wait-listed control children for each of these programs serving as a comparison group receiving "low-quality" care. (Seventeen percent of the families of the 3-year-old cohort control group found a way to enroll their children in Head Start; the comparable figure for the 4-year-old cohort was 14 percent. These will either be deleted from the study or, if a measure of their program quality is available, assigned to the level of quality indicated by this measure.) Although children in the EHS and HSIS were not randomly assigned to these different quality groups, selection problems should be minimal. In the case of EHS, families had little choice between the two types of programs – they would have had to move to a different location to opt for higher quality care. While there may have been more opportunities to shop around in the case of Head Start, we expect those data to show a high degree of geographic clustering on program quality.

To investigate our assumption that the quality of care for the control-group counterparts to children in below-average Head Start quality constitutes "predicted low-quality care, we performed two types of

calculations using ECLS-B data. We first investigated the overall distribution of quality for low-income children in Head Start vs. other care arrangements. We found a mean difference of approximately 0.4, or about 40% of a standard deviation of quality; non-Head Start children receive lower quality care. Second, we found strong geographic clustering of program quality, which will likely be even stronger when we use actual centers as the aggregation unit. This suggests that using the comparison groups in below-median Head Start sites as the groups receiving low-quality care is likely to be successful. With a "low-quality" treatment group thus defined, we will estimate interaction equations such as:

(3) Later Skills = b_0 + b_1 LQT + b_2 Protect + b_3 LQT x Protect + *j* (Controls) + u

In this equation, "LQT" indicates low-quality treatment and "Protect" indicates an index of protective factors. As measures of possibly protective personal and family characteristics we include a high resource and low-risk family environment and having a positive (not-difficult) temperament. Our test of the protective hypothesis is for a positive interaction between low-quality child care and protective family and personal characteristics. Our test of the cumulative disadvantage hypothesis is for a negative interaction between low-quality child care and indicators of high personal and family risk.

C.3.c.4 Aim 4 analyses. Our fourth aim tests the strong predictions coming out of the *differential susceptibility* hypothesis. The hypothesis has two parts – that, for children with difficult temperaments, (a) high-quality child care magnifies positive impacts, and (b) low-quality care magnifies negative impacts. In this case, we need three levels of quality rather than the two typically provided in the treatment vs. control contrasts in our experiments and quasi-experiment. To test the low-quality portion of the differential susceptibility hypothesis, we again turn to the EHS and HSIS studies, and divide the treatment groups into above and below-median groups, with the control group to the below-median treatment centers defined as receiving "low-quality" care. As in the Aim 3 analysis, we distinguish two kinds of quality in the EHS based on implementation and type of program. For HSIS, we distinguish quality using the ECERS. Running a model akin to Equation (3), with "difficult temperament" substituted for "Protect", we expect a negative interaction between difficult temperament and low-quality care. To test for the positive interaction at the positive end of the quality scale, we repeat our procedures in Aims 1 and 2, expecting that the interaction of difficult temperament and high-quality treatment to have a positive and significant impact on child outcomes.

C.3.c.5 Aim 5 analyses. Our fifth aim uses data from the HSIS and PCER studies to identify those detailed program characteristics that raise performance for the most at-risk students. These include classroom characteristics such as program quality measured by the ECERS, teacher sensitivity to students as measured by the Arnett; TBRS measures of the extent and nature of instruction in literacy and math; whether the center uses a curriculum; the child/staff ratio, center size, teacher turnover; and teacher demographic and training characteristics. We will estimate both conventional and instrumental variables versions of these models.

An example of the nonexperimental equations to be estimated for the HSIS analysis is shown in equation (4), which tests whether either the classroom's adult/student ratio and/or preschool teacher's education have larger effects for low SES children than high SES children.

(4) Later Skills = b_0 + b_1 Adult/Student Ratio + b_2 Teacher Education + b_3 At-risk + b_4 At-risk x Adult/Student Ratio + b_5 At-risk x Teacher Education + *j* (Controls) + u

A variety of program characteristics will be tested in this manner. Importantly, this equation will be estimated with fixed effect controls for the HSIS or PCER center to which the child applied and either won or lost the lottery-generated chance of getting in. In (4), the statistical significance of b4 and b5 test whether the program characteristics produce different readiness effects for at-risk versus not at-risk students. Although (4) includes only two process measures, our full analysis will control for all of the process measures in order to avoid attributing to a given process measure what should really be attributed to another.

Estimates of b4 and b5 might be biased by unmeasured variables that influence a family's choice of process characteristics in their chosen child care center as well as their children's outcomes. As discussed by Ludwig and Kling (2007) and Bloom, Zhu and Unlu (2010), we can employ instrumental variables procedures using the random assignment feature of both HSIS and PCER studies to exploit program and site (i.e., classroom) variation in process quality. Specifically, we can use interactions between treatment group assignments (T) and site (S) as instrumental variables to isolate experimentally-induced variations in process variables such as classroom quality. For a single classroom quality indicator (W), and using Controls to denote baseline covariates, we have:

(5) W = T*S $\square_1$ + S $\square_2$ + Controls $\square_1$ + $\square_1$

(6) Later Skills = W $\square_1$ + S $\square_2$ + Controls $\square_2$ + $\square_2$

Interaction between, for example, the child's at-risk status and classroom quality add interaction terms between W and the at-risk indicator. Site dummies in (5) ensure that all of the variation in W comes from within-site treatment-control differences. This ensures that the prediction of Later Skills with W in (6) is identified from exogenous (i.e., random-assignment) variation. Unfortunately, this will still yield biased estimates of $\square_1$ if there are multiple, correlated process mediators. In the case of, say, two correlated mediators W_1 and W_2, equation (5) becomes (5a) and (5b) with the two W components as dependent variables, and equation (6) includes both W_1 and W_2 and their possible interaction with child characteristics. Multiple mediator models are more difficult to identify, but at least in the case of the HSIS data, the many centers, each with different levels of process quality and different patterns of impacts, provide quite a bit of power to estimate such models.

The success of IV models such as (5)-(6) depends on the strength of first stage prediction. Although we are unable to present F-ratios from the yet-to-be released HSIS data, Puma et al. (2010) document very strong effects of the assignment to Head Start on all of our process variables. For example, random assignment of 4-year-olds to the Head Start group increased the percentage with a high-quality classroom environment (5 or above on ECERS/FDCRS) from 38.3% to 71.5%, increased the percentage in a high-quality literacy instruction classroom (7 or above on the TBRS) from 28.1% to 61.5%, and increased the percentage of their teachers with a BA from 19.7% to 30.8%. Similarly, two of the PCER programs – *Bright Beginnings* (BB) and *DLM Early Childhood Express with Open Court* (DLM with OC) – had statistically significant TBRS measures of phonological instruction, with effect sizes of 1.53 and 1.41, respectively (as well as significant effects on student-level outcomes affected by such instruction).

C.3.d Power

C.3.d.1 Power in the HSIS. Puma et al. (2001) calculate likely minimum detectable effects at 80% power for the 3- and 4-year-old subgroups, and subgroups *within* the 3- and 4-year-old samples. For the PPVT, the 3- or 4-year-old samples are each forecast to detect impacts of .14 sd at 80% power. For impact differences between equal-sized subgroups of the 3- or 4-year old samples, the minimum detectable effect increases to .28 sd; a 75/25 subgroup split increases the minimum detectable effect still further to .35 sd. Bloom, Zhu, and Unlu (2010) show that two conditions must be met for the instrumental variables technique to produce unbiased estimates capable of detecting significant effects of program characteristics – the effect of the randomly assigned treatment on the program characteristic must be sufficiently strong, and the overall sample size must be sufficiently large. Fortunately, these conditions appear to be met for the HSIS data, and for two of the PCER substudies – BB and DLM with OC (PCER 2008).

C.3.d.2 Power in the other datasets. EHS enrolled about 2,000 children. Love et al. (2002) provide subgroup estimates for their White (n=769), Black (n=709) and Hispanic (n=475) subsamples. Applying similar logic as with the HSIS, it appears that EHS is able to detect subgroup differences between Whites and Blacks of about .28 sd but for Hispanics vs. either White or Blacks, the minimum detectable effect size difference is closer to .35 sd. The IHDP successfully followed 875 of its sample through age 5. Brooks-Gunn et al. (1994) report subgroup estimates for their heavier (n=295) and lighter (n=580) subgroups. Applying similar logic as with the HSIS, IHDP appears able to detect subgroup differences between lighter and heavier babies of about .40 sd. Case counts (n=3,028) in the Oklahoma Pre-K data are substantial, although the regression discontinuity design provides less power than would a RCT. Gormley et al. (2005) provide subgroup estimates for their White (n=925), Black (n=969) and Hispanic (n=321) subsamples. Applying similar logic as with the HSIS and taking their Applied Problem scores as an example, it appears that OK Pre-K is able to detect subgroup differences between Whites and Blacks of about .40 sd, but for Hispanics vs. either White or Blacks, the minimum detectable effect size difference is closer to .60 sd. The PCER sample size of 2,911 is also substantial, and should provide adequate power to detect effects in the range above. To summarize, the rank order of power to detect subgroup difference is HSIS (most power), EHS, IHDP, OK Pre-K, and PCER. Fortunately, the two studies with the most power, the HSIS and EHS, are collectively involved in testing all of our hypotheses.

C.3.e Timeline and Organization of the Project

Addressing five research aims with comparable analyses of five datasets is challenging. However, we believe that it can be successfully accomplished because our researchers are experienced with some of the data bases, and all but one (Sojourner) are resident at UC Irvine. We will operate as a research collaborative, with all

four PIs, plus one graduate assistant throughout the project, and the other part-time, meeting on a regular basis. (Sojourner will participate by telephone.) Farkas will be the overall coordinator, with responsibility for the quality and timeliness of all research products, but each of the PIs will play a major role in overseeing the analyses of the data bases for which he or she has taken responsibility: IHDP – Duncan/Sojourner; EHS - Burchinal; HSIS - Farkas; Oklahoma – Duncan; PCER – Farkas/Burchinal. Assuming a start date of July 1, 2011, the approximate timeline for completion of the tasks will be as follows (separately by months):

1-12 Preparation of variables and descriptive statistics for all data bases. (This will be facilitated by the fact that in our work on our IES grant, we have already been analyzing the IHDP and EHS data bases.)

13-24 Analysis of data from HSIS and PCER to address Aims 1 – 5. Begin presentations on results. Interactions between the personnel on this subproject and the Irvine Network will suggest additional analyses.

25-36 Analysis of data from IHDP and EHS to address Aims 1 – 5. Continue presentations and consultation with the Irvine Network; begin to prepare papers on findings.

37-48 Analysis of data from Oklahoma to address Aims 1 – 5. Continue interaction with Irvine Network, presentations, and preparation of papers.

49-60 Consistency checks of results across all data bases. Results combined in statistical meta-analyses. Continue interaction with Irvine Network, presentations, and preparation of papers.

C.3.f Preliminary Studies

The research team, led by George Farkas, has recently completed its first year of work on a two-year research project (begun July 1, 2009) funded by the IES, U.S. Department of Education, "Preschool Program Impacts on School Readiness: Variation by Prior Child Language and Attention Skills, and the Quality of Infant/Toddler Care." This project involves the analysis of six datasets, three of which – data from the IHDP, EHS, and HSIS – will also be analyzed in the present study. Thus, we have already made a beginning working together on a subset of the datasets to be analyzed in this project. The proposed project involves different analyses of these data than is called for in the IES project, as well as analysis of two – the Oklahoma and PCER -- not analyzed in the IES project. However, work on the proposed project will build on the knowledge and experience gained from the IES project. In particular, the IES project tests whether high-quality child care has different effects for children of differing cognitive and attention skills, and does so using six databases, three of which are experimental. By contrast, this project tests five different *theories* of the consequences of the match between child/family characteristics and the quality of child care (both high and low quality), estimates differential child care effects across population subgroups defined by family environment (SES and family risk factors), baseline cognitive skills, age and temperament, and does so using four experimental and one quasi-experimental database. By extending the work on the IES project, and using experimental and quasi-experimental databases, our project provides a comprehensive assessment of the ability of high-quality child care to reduce achievement gaps at kindergarten entry for varying population subgroups, as well as of the effects of particularly low-quality child care on these outcomes for these subgroups.

C.3.g Interdependence with the Core and Other Subprojects

Our subproject complements the others by focusing on a distinct childhood period – from birth to school entry. It employs the child-policy match framework and is interdisciplinary, involving a number of Network members (Farkas, Duncan, Burchinal and Vandell) in providing theoretical perspectives relevant to Project II's focus on child/health curriculum matches in middle and high school and Project III's focus on the distribution of outcomes from primary and middle-school interventions. The closest link is to Project III's examination of distributional impacts from the HSIS. Our project employs a conventional baseline interaction approach while Project III develops a new non-parametric approach for understanding similar issues. We will work closely with the Project III project team (Burchinal is in both groups) to understand commonalities and differences in results. Finally, our achievement and behavioral outcomes are key independent variables for Project IV's analyses of the role of middle childhood skills and behavior for adult outcomes. Project IV will thus provide guidance as to which of the outcomes we measure matter the most for children's long-run success. This subproject also draws heavily on the intervention and statistical expertise of External Advisory Committee members. Ludwig, Bloom and Reardon are experts on the instrumental variables procedure we propose. Farran and Clement bring experience in early childhood intervention, including the PCER data employed in our subproject, while Dodge has developed interventions spanning childhood and adolescence.