

Recommendations

OCDEL Short-Term Efficacy Evaluation of State-Supported Early Learning Programs

Provided by
The Universities Children's Policy Collaborative¹

General Background

In March, 2007, the Office of Child Development and Early Learning (OCDEL) in Pennsylvania's Departments of Education and Public Welfare commissioned the Universities Children's Policy Collaborative (UCPC) to conduct background research and create a set of recommendations for OCDEL to subsequently commission a short-term efficacy evaluation of state-supported early learning programs in Pennsylvania. UCPC is a collaborative group of representatives from the Prevention Research Center at Penn State University, the University of Pittsburgh Office of Child Development, and Temple University, all state-related institutions.

This document presents the pros and cons of various choices regarding programs, evaluation designs, sampling, measurements, and analyses, made in collaboration with OCDEL, that lead to a proposed set of guidelines in conducting such an evaluation under independent and separate contractual arrangements, tentatively beginning in January, 2008. However, it must be recognized that once the actual study is begun new circumstances inevitably will be discovered that will markedly limit or expand the proposed plan. Thus, we present recommendations, not prescriptions, that can and should be changed in the face of unforeseen political, financial, and practical circumstances existing at the time such a project is actually implemented.

Project Overview

Purpose. The purpose of the proposed evaluation is to determine 1) the level of quality of classrooms for 3-5-year-olds in several state supported programs and to

¹ The Universities Children's Policy Collaborative (UCPC) is a partnership of the University of Pittsburgh Office of Child Development, the Prevention Research Center at Pennsylvania State University, and faculty at Temple University dedicated to bringing research and best practice information to the policy process in Pennsylvania to contribute to the health, education, and welfare of the Commonwealth's children, youth, and families. Principal contributors to this report are Robert B. McCall and Christina J. Groark, University of Pittsburgh Office of Child Development; Richard J. Fiene, Prevention Research Center at Pennsylvania State University; and Marsha Weinraub, Temple University. UCPC appreciates the advice of several consultants including Heather Bachman, Assistant Professor of Education, University of Pittsburgh; Steven Bagnato, Professor of Pediatrics, Director of the Early Childhood Partnerships, University of Pittsburgh; JeeWon Cheong, Assistant Professor of Psychology, University of Pittsburgh; David Johnson, Professor of Sociology, Human Development and Family Studies, and Demography, Pennsylvania State University; Kathleen McCartney, Dean, College of Education, Harvard University; Samuel J. Meisels, President, Erikson Institute; Elizabeth Votruba-Drzal, Assistant Professor of Psychology, University of Pittsburgh.

compare those levels with classrooms in non-supported child care from 2002 and the present, 2) if children enrolled in each of the several state-supported early childhood care and education programs make more developmental progress than children not enrolled in those programs, and 3) whether children benefit relatively more or less as a function of type of program, administrative characteristics of the program, and certain characteristics of the child and his or her family to the extent possible.

- *Type of Program* – The primary *state-supported* programs to be assessed are Keystone STARS, Head Start, Pre-K Counts (Accountability Block Grant Pre-K), and segregated Early Intervention. These programs will be compared to *non-supported child care* programs (“Non-Supported CC”). Children with disabilities attending segregated state-supported early intervention programs will be compared with children with disabilities in integrated settings as sampled within the other types of supported programs to the extent they are present in these programs. Comparisons will also be made for center, group, and family sites where these types of sites are available within program type.
- *Developmental progress* – The evaluation will assess the general developmental progress of children 36-60 mos. of age in these programs with assessments that reflect their general developmental progress, especially aspects of development that are related to the skills widely thought to contribute to school readiness (e.g., vocabulary, language, basic concepts, preliteracy, prenumeracy) and are related to the Pennsylvania State Pre-K Educational Standards. No attempt will be made to assess child characteristics that are more specifically related to the particular aims and activities of the program type, because program-specific evaluations are being conducted separately for each type of program.
- *Comparison groups* – “Children not enrolled in state-supported programs” will be defined in several ways.
 - *Classroom quality.* Classroom quality will be compared to two standards. First non-supported child care sampled in the present study represent contemporary non-supported care. However, such sites are those who have not joined Keystone STARS; so while they are indeed current non-supported programs, they probably reflect some selection bias. So, classroom quality as assessed by Fiene et al. (2002) in 1996 and 2002, before the advent of Keystone STARS will also be used as a non-supported benchmark.
 - *Children’s development.* First, a sample of children currently enrolled in regulated child care programs in Pennsylvania that are not part of Keystone STARS and do not receive state support of any kind (ignoring any state support to individual children/families) will constitute the primary comparison group representing “non-state supported programs.”

- Second, national samples of children that constitute the standardization samples for the measuring instruments that provide standardized scores or percentile ranks for individual children in the current project will collectively represent children exposed to “a reasonably representative sample of all types of services available across the USA in the year of the standardization study.”
- Third, the pre-test scores of children actually assessed in the proposed study can be used to predict what level of test performance reasonably could be expected from them at the age of post-testing if they had not been exposed to the state-supported program. This constructed comparison group is most closely matched to the participating children (i.e., it is based on the pretest scores of the same children) and the most sensitive comparison to determine the developmental progress of different groups of children.
- *Program quality.* Program quality consists of the general level of environmental supports for learning and the quality of the social-emotional interaction between teacher and children. In addition, to observational measures of classroom quality, a variety of other indices of quality will be obtained (e.g., class size, teacher-child ratio, use of curriculum, teacher education and experience, etc.). These measures of quality will be both outcomes that reflect the quality of various Pennsylvania programs, but they can be examined as mediators of program effects on children’s development (i.e., what is the relation of quality within a program to child outcomes; after quality of programs is considered, does the particular program relate to children’s development?).
- *Program characteristics* – To the extent practical, a variety of program characteristics and administrative arrangements that are unevenly distributed across the main program types listed above will be assessed to determine if children develop differently as a function of those characteristics, including 1) the nature of the administrative unit operating the sites; 2) urban versus rural location; 3) family, group home, and center sites; and 4) education/income level of families participating in each site (if possible). Within Keystone STARS programs, children’s developmental progress will be examined as a function of whether the program had 1, 2, 3, or 4 STARS.
- *Child characteristics* – To the extent practical, the project should attempt to assess whether children develop differently across and within program type as a function of 1) family education/income; 2) language used in the home; 3) predominant race/ethnicity; and 4) children with disabilities in segregated versus integrated settings.
- *Children and program exposure* – The proposed study will focus on children 36-60 mos. of age, who will be given a pre-test in the early fall and a post-test in late spring of the following year. Thus, the study will examine the developmental

benefits of approximately 7-8 mos. of exposure to the program. To the extent practical, longer exposures to the program up to approximately 20-24 mos. also will be examined.

Questions to be addressed. The proposed project will attempt to address several major questions (and a variety of more specific questions).

- Do children enrolled in state-supported programs display greater developmental progress during an academic year of enrollment than children enrolled in programs that are not state supported?
- Do children display greater developmental progress, apart from the level of development at which they entered the program, as a function of whether the program is Keystone STARS, Head Start, or Pre-K Counts (ABG Pre-K)? Do Children display different rates of development as a function of whether the child care program site is a center, group, or family environment, and whether each of these program types separately is associated with better child developmental progress than children enrolled in non-supported child care programs?
- To the extent practical, do children show relatively greater developmental progress if they are enrolled in programs with certain structural or administrative characteristics (e.g., programs that are combinations of more than one program type, urban versus rural, income level of participating families) across different program types as well as within different program types, although the ability to answer these questions will depend on the extent to which programs with each characteristic are sufficiently and fairly represented within each major program types?
- Do children display greater developmental progress as a function of family education/income level, language in the home, and racial/ethnicity across program type and within program type and administrative characteristics, although the ability to answer these questions also depends on the extent and distribution of child/family characteristics across and within program type and characteristic?
- Do children with disabilities do better in segregated centers or integrated in other supported sites? The ability to answer this question will depend on the number of children with disabilities enrolled in integrated settings.
- Do children display greater developmental progress as a function of more exposure to the program?

OCDEL Short-Term Efficacy Evaluation of State-Supported Early Learning Programs

Provided by
The Universities Children's Policy Collaborative

Proposed Project Guidelines

The pros and cons of various choices and the rationale for the proposed guidelines is presented below with respect to the programs to be studied, sampling procedures, measurements, assessment procedures, and data analyses. These guidelines will be modified by the contractee in collaboration with OCDEL during the course of implementing the proposed project as unanticipated circumstances and opportunities arise.

State-Supported Programs to be Evaluated

The primary programs to be evaluated are Keystone STARS (STAR levels 1, 2, and 3+4), state-supported Head Start programs, Pre-K Counts/ABG Pre-K, segregated Early Intervention programs, and non-state supported child care. However, these programs are not necessarily mutually exclusive; and while a site may be sampled as a Pre-K site it may also be in Keystone STARS, for example.

Program Descriptions

Briefly, the major program types are as follows:

- **Keystone STARS.** Child care *centers*, *group homes* (6-12 children), and *family care* (6 or fewer children) can apply to the Commonwealth to be graded on quality, and sites at certain quality levels (called STAR Levels) receive financial and technical support to help them upgrade to a higher level. Sites are designated as “Start with STARS,” STAR 1, 2, 3, or 4 based on meeting certain steps along several dimensions of quality (including the ECERS Scale). Sites already nationally accredited by NAEYC, for example, are automatically given STAR 4 designations.
- **State-Supported Head Start.** These are Head Start programs for low-income children that receive state financial support to expand enrollment or extend their days/hours of services. In addition to *center-based* services, Head Start can be offered as a *home-based* program consisting of weekly home visits to promote parenting and referrals to health and family services plus regular opportunities to visit centers for parental and child enrichment activities.
- **Pre-K Counts/Accountability Block Grant Pre-K.** The Accountability Block Grant (ABG) Pre-K program offers school districts funds for additional

programming in several areas including quality Pre-K services. The Pre-K Counts program, proposed for implementation in pending legislation, provides *Head Start, school districts, licensed nursery schools, and child care centers and group homes* with additional funding to expand enrollment and extend the daily length of service. Programs must meet a list of quality requirements, and services are targeted at low-income children. The ABG and Pre-K Counts programs are likely to be combined and expanded (if approved) in the near future, and will be called Pre-K Counts in this document.

- **Early Intervention.** State-supported segregated services in *centers* for children with disabilities are operated by school districts and Intermediate Units. Children with disabilities served in integrated settings in any of the other program types will be sampled as available.
- **Non-state supported child care.** These are *center, group homes, and family child care* that are not state supported and not in Keystone STARS. They constitute the current “non-state-supported services” that is the primary comparison group for state-supported program types.

Notice that several characteristics of programs will be embedded in the four major program types, although they will be unevenly represented (in some cases, not represented at all) in each program type, including programs in urban versus rural locations, income level of participants; whether the program is categorized as a family care, group home, or center; and the type of administrative unit especially within Pre-K Counts (i.e., Head Start, school district, licensed nursery schools, child care).

Rationale for Not Testing Children Under 36 Months

The proposed evaluation guidelines suggest limiting assessments to children 36-60 mos. of age and not assessing children birth to 35 mos. There are several reasons for this suggestion, especially for omitting children birth to 35 mos. First, program characteristics change dramatically for children younger versus older than 36 mos., because state regulations pertaining to staff-child ratios, group sizes, space, and other program characteristics are different for children younger than 36 mos. than children who are older than 36 mos.

Second, assessments on infants and toddlers are less reliable (at least within the first 6 mos. of life; McCall, 1979), and the skills represented on tests of general development during the first 18 mos. of life are very different from those represented on assessments aimed at 3-5 yr. olds even though a single test covers this entire age range (even assessments up to 30 mos. of age are very different from those at 36-60 mos. except for the prevalence of elementary vocabulary items; McCall, Eichorn, & Hogarty, 1977).

Third, research indicates that differences in home and out-of-home environments are only minimally related to the test performance of children in the first 2-3 years of life

(McCall, 1979), which suggests that finding differences in children's development as a function of program type is much less likely among children birth-35 mos.

Finally, the literature indicates very little predictability from general developmental assessments made birth to 35 mos. to similar developmental assessments made between 36 and 60 mos. and almost no predictability to school performance or general developmental performance in children 6 yrs. of age or older (Mehaffie & McCall, 2002).

Thus, it seems substantially less fruitful to evaluate children birth-35 mos. of age in different state-supported programs than children 36-60 mos. who are more likely to be influenced by program type and whose assessments are more likely to have face validity as measures of school readiness and be somewhat more likely to predict early school success.

The decision not to assess children birth-35 mos. means that the Nurse-Family Partnership for children birth to 2 yrs., the Parent-Child Home Program for children 1.5 to 3 yrs. of age, and children birth – 35 mos. in state-supported early intervention would be excluded from the proposed evaluation project. Their omission from this evaluation is compensated to some extent by the fact that most of these programs have fairly substantial evaluations designed specifically to evaluate the quality of their programs and its effects on children with measures more specifically related to their program aims and activities than the current project's assessments of general child development in infants and toddlers.

Rationale for Not Sampling Other Types of Programs

In an ideal evaluation, the different types of programs to be compared are independent of each other and mutually exclusive, but this is not the case to varying extents for state-supported programs. Indeed, even the state-supported programs selected as primary sampling units (Keystone STARS, Head Start, Pre-K Counts, Early Intervention) are not independent and mutually exclusive because some programs can qualify as being in more than one of these program types. Further, other program characteristics (e.g., family/group/center, urban versus rural) and child/family characteristics (e.g., education/income, disability, English home language, etc.) will also be nested with program type.

It is proposed that these nested characteristics not constitute groups that would be deliberately sampled but would occur in approximately their representative frequencies within samples of the other three program types. The rationale for this suggestion is that the project would become exceedingly large and nearly impractical to conduct within a limited period of time if programs and children with all three characteristics were deliberately sampled (but see below). These characteristics will be related to child outcome to the extent that the characteristics are represented in the sampling.

Sampling Strategy

Population. Table 1 presents in italics the estimated population of sites (not numbers of classrooms or children) as of 2006-2007 (numbers are expected to increase perhaps by 10% or more by the time the proposed study is conducted) for some of the four major program types to be deliberately sampled. Table 1 breaks down the major program types into family, group, and center sites where they exist. Note that the number of group sites is small, especially within higher Keystone STARS levels, so they may need to be combined with centers for some analyses. Also, Home-Based Head Start programs are unique and are neither group nor family sites. The table provides the relative numbers of sites in each of the four sampling groups as well as in some major subgroups within each program type (e.g., family, group, center; Keystone STAR levels), but such estimates cannot be accurately made at this time for Home-Based Head Start or for the emerging Pre-K Counts program.

Sampling Steps – Site Selection

A suggested Step 1 is to randomly select a certain number of sites within each of the four major program types and family/group/center types. While some types and subgroups of programs are relatively infrequent in the population (see Table 1), the sampling strategy would pick similar numbers of sites in each program type to produce reasonably sized samples of all major categories. This means program types and the total sample will not proportionately represent the frequency of program types (or children) in the Commonwealth, but this can be estimated later by extrapolating results for each type in proportion to its frequency in the state. Instead, samples of each program type and characteristic will be more likely to be of sufficient size to reveal differences between types and characteristics if such actually exist (i.e., afford sufficient “statistical power”) and produce reliable results that are less influenced by a very few, perhaps extreme sites. The initial sample should be at least 50%-75% larger than is likely needed depending on type of program (Fiene et al., 2002) to cover refusals to participate.

Step 2 is to continue to sample with replacement (i.e., randomly select a new site from the population and randomly delete sites already selected from an over-represented type) so that the selected sites are clustered geographically to minimize the travel of assessors. A truly stratified random sample would likely be all over the state and represent substantial logistical burdens to assessors who must assess all children within approximately 4-5 weeks in the fall and again in 4-5 weeks in the spring. To do so, they cannot be spending a substantial portion of this time traveling long distances to sample one or two sites.

Sampling Steps—Individual Children

The program site is the primary unit so the number of classrooms and children within classrooms actually assessed should be similar at least within major program types and characteristics so large sites do not contribute will be assessed in the first year and Wave II in the second year of the project.

Table 1. Site Population and Minimum Site and Children Samples

State-Supported Keystone STARS

	Center			Group			Family		
	Pop. Sites	Sample Sites Children		Pop; Sites	Sample Sites Children		Pop. Sites	Sample Sites Children	
STAR 1	1160	15	52	212	15	52	408	24	48
STAR 2	503	15	52	44	15	52	50	24	48
STAR 3	151	15	52	5	5	20	10	10	20
STAR 4¹	366	15	52	17	15	52	47	24	48
	2180	60	208	278	50	176	515	82	164

Non-Supported Child Care

	1283	51	178	387	51	178	2972	84	168
--	------	----	-----	-----	----	-----	------	----	-----

Head Start

	Center			Home-Based ³		
	Pop. Sites	Sample Sites Children		Pop. Sites	Sample Sites Children	
	56	30	105	?	30?	50?

Pre-K Counts²

	Center			Group		
	Pop. Sites	Sample Sites Children		Pop. Sites	Sample Sites Children	
Keystone STARS	123	15	52	17	17	59
Head Start	13	13	45	X	X	X
Lic. Nursery Sch.	41	15	52	X	X	X
School Districts	43	15	52	X	X	X
	220	58	201	17	17	59

Early Intervention

	Center			Group			Families		
	Pop. Sites	Sample Sites Children		Pop. Sites	Sample Sites Children		Pop. Sites	Sample Sites Children	
Segregated:	31 <i>locations</i> 180-200 <i>classes</i>	30	105	X	X	X	X	X	X
Integrated (Supported)⁴	?	?	96?	?	?	25?	?	?	20?
	31	30	201	?	?	25?	?	?	20?

¹Includes sites accredited by professional organizations.

²Pre-K Cts/ABG Pre-K, abbreviated Pre-K Cts. in this document, is an emerging combination of two current programs: 1) Pre-K Counts, a public-private program of support to expand early care and education, and 2) Accountability Block Grants for Pre-K in which school districts may apply and receive funds to support pre-K programs that adhere to state standards of quality. The number of such sites likely present in 2008 is uncertain and dependent on new appropriations. The figures in this table represent the number of sites that have submitted letters of intent as of May 2007 in advance of passage of the funding legislation.

³Home-Based Head Start is neither a Group Home nor Family Care service. Instead, a parent receives weekly home visits that promote effective parenting and early education plus referrals to health care and family services. Parent and child also attend center-based activities to provide the child with social experiences and parents with additional training and recreation.

⁴All children with disabilities who are integrated into other supported (i.e., Keystone STARS, Head Start, Pre-K Counts) will be specifically sampled in addition to the other sampling of children. The number of cases actually available is uncertain.

X - No meaningful number of cases exist.

? – Population and sample cannot be accurately estimated.

disproportionately to the results for that program type. One could argue that sites should be represented by children in proportion to their enrollment—assess a fixed percentage, not number, of children within a site—to produce results that reflect the population of children. We suspect OCDEL, as a government agency supporting program types and sites within programs, is more interested in the program type and site as units of inquiry, which would be fairly represented by assessing the same number of classrooms and children in each site.

So at the site level, Step 3 is to select one or two classrooms or groups per site that have the most children meeting preferential criteria, which are mainly 3-year-olds but also some four-year-olds both new to the program and some with a year or more experience) and full-time attendees (except if sampling will be very large). Within classroom/group, select up to 6 children. The target is to assess 4-5 children per center, 3-4 per group and 1-2 per family site; overselection is necessary to cover refusals or the inability to obtain informed consent, although all children whose parents give consent should be assessed.

Note that children who are three years old and full-time attendees are favored. Such children are more likely to show greater improvements in their 8-9 mos. of program participation than children who are part-time or in their second year. However, children who experience the program for more than 8-9 mos. are likely to show a cumulative benefit over their longer residency (up to 20 mos.); so some four-year-olds with and without previous experience with the program will also be selected (see analysis section). Note that it is desirable to have children who are new to the program who span the age range of 36-50+ months at pre-test so we can use their pretest scores to estimate expected scores across age for inexperienced children in each program type and income level as an important comparison condition (see below). We recognize it will be difficult to balance perfectly experience across age and will test and adjust for certain biases.

Step 4 might include over-sampling certain types of individual children within a site. For example, one might deliberately test all children in a supported site who have a disability as defined by an IEP, non-English speaking home, low family income, and minority race/ethnicity. These over-samples are in addition to the initially selected

children and are so labeled in the database. If a child representing these underrepresented characteristics is selected as part of the initial sampling, he or she is retained in the initial sample to maintain its representativeness.

Minimum Sample

Table 1 also presents a minimum sample of sites and children (in Roman type) for each of the sampling units. An effort was made to create a sample that was approximately equal sized for each of the major sampling units (rather than proportional to either the number of sites or children), because OCDEL's interest is in comparing different program types and certain characteristics within programs rather than to estimate the overall quality of care in the state. The sampling numbers represent the desired minimum sample having complete data, so the initial pre-testing sample will need to be a bit bigger to anticipate attrition of children who leave a program between fall and spring. The sample assumes approximately 3.5 children will have complete data from each center or group home, but 2 children per family site; 4 children may be assessed per classroom and site for those types of programs with small populations. For some program types and subcategories, the size of the population is not known precisely, so neither is the sample known precisely until sampling actually begins. Further, there is no way of knowing how many children with disabilities are in the various integrated settings, so an approximately number has been estimated based on an arbitrary estimate that 5% of the children will have a disability.

The sample presented in Table 1 is considered to be the smallest needed for reasonable accuracy and power to detect differences between the major sampling groups. However, the more sites and children assessed the better the project can assess differences between programs and children on other characteristics that vary within and between programs, such as full-day versus half-day attendance (the minimum sample favors full-day children, but an expanded sample could increase the number of half-day children), children who have more experience with the program than three-year-olds and new four-year-olds (i.e., four-year-olds who have been in the program for at least a year prior to this study), children with disabilities in integrated settings, as well as various child/family characteristics. One way to increase sample size is to conduct two Waves of data collection, one in the first year and one in the second year of a longer project.

Note in this sample scheme that Keystone STARS and the non-supported child care that is the base comparison for all program types are the most heavily sampled, because the difference between STAR levels and center/group/family type sites is desired by OCDEL. Similarly, Pre-K Counts has more sites and children to be sampled than Head Start, because OCDEL is interested in the distinction between the administrative units (Keystone STARS, Head Start, licensed nursery schools, school districts) that operate Pre-K Counts.

Measurement Instruments

Children's Assessment

Criteria for the children's assessment instrument. The assessment instrument should reflect "general behavioral development" with an emphasis on skills likely to be taught to varying extents by the major program types that contribute to school readiness and that represent the State Education Standards for Pre-K. Ideally the assessment tool should:

- Be a single instrument or set of subtests to be used for all children, with and without disabilities, across all types of program types.
- Have a balanced content to reflect a blend of practical skills likely to be taught in early care and education programs plus conceptual/abstract abilities that are reflective of skills we would expect children 36-60 mos. of age to acquire in early care and education programs. These skills should contribute to school readiness (e.g., vocabulary, language fluency; emergent literacy; emergent numeracy; knowledge of fundamental concepts of colors, shapes, etc.; basic concepts; reasoning skills), and forecast at least to some extent early achievement in school.
- Reflect to the extent possible those Pennsylvania Early Learning Standards and the basic principles of the three child/teaching/assessment strategies emphasized by OCDEL for early care and education programs (i.e., Work Sampling System and the Ounce Scale).
- Have national norms that provide standardized scores or percentile ranks for total scores as well as subtest scores so that each child will receive an age-invariant score on his or her pre-test and post-test that is a meaningful comparison with national norms regardless of the child's age or time in the program.
- Be administered to each child in a relatively short period of time, such as 20-30 minutes, because a substantial number of children must be assessed in a 4-5 week period and because young children will not sit still and concentrate for longer periods of time.
- Be appropriate for children 36-60 mos. of age and not unfairly penalize children with disabilities, children whose primary language in the home is not English, and children of different racial/ethnic origins. We recognize that such factors necessarily influence children's assessment performance, but this "bias" also exists in early care programs and early school performance.
- Have good psychometric information on reliability, validity, and relation to other commonly used but more extensive assessments of general development and with early school performance.

Authentic/teacher/parent/criterion-referenced assessments versus independent examiner/norm-referenced assessments. Pennsylvania has encouraged its early care and education programs to use assessment and curriculum systems within its early care and education programs that are criterion referenced, which means that a child's progress is charted against his or her own performance on skills in various domains that are known to develop in a given sequence. Further, such assessments are conducted by the teacher (in some cases, the parent), and the assessment system in turn provides the teacher with suggestions for the next skill to be worked on for each individual child. Such systems are said to use criterion-referenced tests that are more "authentic" than other assessments, because the items on the test are precisely those that are taught in the program and because the teacher assesses these skills by observing the child's naturalistic behaviors in the classroom environment. Such assessments have a great deal to recommend them as part of an assessment/curriculum system, but they have certain limitations for the current purpose of assessing program efficacy, especially comparison between different programs, program characteristics, child characteristics, and placing such developmental change in a national context. Specifically:

- *No national norms* – None of the three assessment/curriculum systems advocated by the state has standardization data that empirically support the average age at which various skills should typically be expected in an unselected national sample (even though the skills are categorized by "expected age"). Thus, an age-invariant score cannot be obtained for an individual child that would permit comparing children who start and finish their program residency at different ages; the development during residency of children who begin the program at different levels cannot be easily compared; and the results could not be compared with how well Pennsylvania children in different programs are developing with respect to national averages.
- *Administration by teachers* – The assessment/curriculum versions of child progress are assessed by teachers (and sometimes parents) who rely on their experience with the child in the classroom (i.e., "authentic assessment"). This requires the teacher to have some degree of familiarity with the child's behavior in the classroom and with the assessment system and concepts to be rated before being able to accurately and comprehensively reflect the child's abilities on targeted skills. Unfortunately, in the proposed project, all children must be assessed in the first 4-5 weeks after fall enrollment (and again the last 4-5 weeks in late spring). The longer one gives the teachers the opportunity to observe the child in the fall, the shorter the length of time between pre- and post-test and thus the more difficult it will be to demonstrate developmental improvement in children. Conversely, the less time a teacher is given to observe children at the pre-test, the more likely children will be under-evaluated on the pre-test relative to teacher knowledge of the child at post-test, and developmental gains could be attributed to increased teacher knowledge of the children rather than improvements instilled by the program per se.

- *Teaching for the test* – A major advantage of the assessment/curriculum procedures is that they integrate assessment with the curriculum and deliberately instruct teachers to teach the next skill in the sequence of skills on the assessment. While there is a great deal of similarity, some differences between which assessment/curriculum system an early care and education site uses (or none at all) may mean that assessments made in the context of the curriculum are tied to that curriculum and are not comparable across these curricula.
- *Teachers adjust to the level of the children* – Some assessments within the assessment/curriculum ask the teacher to rate a child’s developmental level “relative to his or her peers,” and such a rating appears to give each child a “score” that is age invariant and could be compared across curricula and program types and characteristics. However, many assessment professionals recognize that the “peers” that a teacher uses as the criterion for such a rating are the children that the teacher sees in the child care site. Therefore, if the site caters to at-risk children who generally display lower levels of developmental performance, that teacher may rate an individual child as doing well relative to these peers but the same child with the same performance would be rated as performing below the level of peers by a teacher working in another site catering to more advantaged children. Thus, such ratings tend not to be comparable across sites in which the level of child performance is different.

Norm referenced tests, which are recommended here, deal with these limitations because they:

- *Provide age invariant scores based on national samples of children.* Such tests provide each child at each assessment with a standardized score or a percentile rank relative to a national sample of children from all kinds of early environments that allow scores to be compared from pre-test to post-test and across sites of various types and characteristics.
- *Are administered by independent trained assessors.* The tests are administered by independent trained assessors who have more specific criteria for judging children’s performance and are not influenced by their experience with a specific program. As a result, they produce assessment results that can be readily compared across sites and program and child characteristics.
- *Can be administered at the beginning of the school year.* Independent assessors can start administering tests shortly after children enroll and adjust to the care environment and do not require several weeks in which to observe the child in the classroom to make the assessments.
- *Use the same assessment items for all programs.* In contrast to the assessment/curriculum strategies which differ from one version to another, a common set of items characterizes the proposed norm-referenced tests, making

the assessment comparable from program to program, although some programs may teach skills that are closer to the assessment content than other programs.

Of course, norm-referenced tests have their own limitations that advocates for authentic assessments in assessment/curricula systems have long identified (see above), namely the child is assessed in a relatively unfamiliar environment by an unfamiliar adult, and the assessment instrument may have items and skills that children do not typically perform (although we will try to minimize this). These circumstances may cause certain children to underperform relative to their actual abilities. However, all children in all programs as well as the standardization sample are assessed in the same way, so no program type should be at a particular disadvantage because of this limitation. Norm-referenced tests generally cannot easily reflect children's social-emotional development without asking the teacher or parent.

Examples of specific assessments. Appendix I presents summaries of several possible general developmental assessment instruments which are examples of many different tests within this broad category.

For example, the Developmental Observation Checklist System (DOCS) is an example of a teacher- or parent-reported assessment that is tied to curriculum, but it also has a standardization sample (1990). The Ages and Stages Questionnaires, which the state has encouraged some early care and education providers to use, is also a teacher- or parent-reported assessment with a standardization sample (somewhat out of date). The CIRCLE assessment allows teachers to rate children using hand-held automated electronic equipment, but it demands children respond to some items in only a few seconds, emphasizing speed rather than knowledge. These assessments have the advantage of using items that are practical in character and thus represent skills that are likely to be taught in early care and education programs, but they have the disadvantage for the current project of being administered by teachers (or parents).

The Battelle Developmental Inventory is one of many tests that are examiner administered and reflect a comprehensive range of early child developmental domains. The Battelle is particularly suitable to children with disabilities, it has a very recent standardization sample, and it contains a mix of items assessing practical skills and more conceptual and abstract thought. The Woodcock Johnson III test has a large set of subscales testing abstract and conceptual thought on the one hand and another large set of subscales that reflect more practical skills likely to be taught in early care and education. Both the Battelle and Woodcock Johnson are very comprehensive, although in different ways, but they can take an hour or more to administer completely, which is likely longer than many young children will tolerate or that project personnel can take to test a single child. Thus, a subset of subscales would need to be selected if one of these instruments were used.

Finally, the Bracken Basic Concepts Scale assesses very practical skills of young children likely to be taught in early care and education programs, it is sensitive to environmental (i.e., program) experiences, it has a subset of subscales collectively called

a School Readiness Composite that only takes 15-20 minutes to administer, these subscales are consistent with the State Education Standards for Pre-K, the norms are based on a 2006 sample, it can be administered by adults after minimum training, it correlates moderately well with other measures including IQ tests, and it relates reasonably well to early school performance.

Of course, there are numerous widely used and well-documented tests of specific early learning skills, such as picture vocabulary, emerging literacy, language development, phonemic awareness and discrimination, and various elements of abstract/conceptual thinking and reasoning. By themselves, they are good tests but quite specific; a set of these designed to be more comprehensive becomes too lengthy to administer in a practical period of time, the costs of the tests and training examiners increases, and the standardization samples differ between tests which makes results for one vs. another at least somewhat incomparable.

Assessment of Program Quality

A major purpose of this project is to assess the quality of programs being supported by the Commonwealth and to make comparisons between the quality of different program types and between each type of program and non-state supported programs.

ECERS/FDCERS. Perhaps the most widely used assessment of early care and educational program environments are the Early Childhood Environmental Rating Scales (ECERS) for centers and group homes and the Family Child Care Environmental Rating Scale (FDCERS) (Harms, Clifford, & Cryer, 2005), and indeed the Commonwealth of Pennsylvania uses these rating scales as one of the criteria in allocating different STAR levels within Keystone STARS. Should this project use the same rating scale as used in Keystone STARS or should it use a different scale to “cross validate” the Keystone STAR rankings on a separate measure?

HOME. Possible alternatives to the ECERS and FDCERS are the HOME Inventory (Caldwell & Bradley, 1984), originally developed to assess the parent and home environment of a child but which recently has been adapted to be used for group care settings. It has the advantage of having more items on caregiver responsivity and caregiver acceptance of children than do the ECERS and FDCERS, which tend to emphasize the physical and organizational aspects of the environment to a greater extent than teacher-child interactions. Disadvantages of the HOME are that it is less widely used in group settings, there are fewer national databases that have used the HOME (the NICHD Early Childhood Network used it), and items are simply scored yes/no and therefore do not provide as fine-grained assessment of various characteristics.

Advantages to ECERS/FDCERS. The advantages of using the ECERS and FDCERS are that they are the most widely used such assessments in the country, OCDEL has decided that they are major criteria of quality in early care and education for the

Commonwealth, and data exist before the state began to support programs that can be used to compare current program quality.

Teacher-child interactions. Many scholars and practitioners believe that the ECERS, FDCERS, and HOME Inventory do not emphasize enough the nature and extent of the teacher-child interactions, especially their social and emotional characteristics. These professionals suggest that the child's development may be much more closely associated with the nature and extent of these interactions and the relationship developed between teacher and child than with the nature of the physical and organizational environment that are emphasized on these environmental rating scales. The Caregiver Interaction Scale (CIS) (Arnett, 1989) deliberately assesses teacher-child interactions, and represents a possible additional assessment that would complement the environmental rating scales. However, the CIS seems to provide high scores for too many teachers, and it is also somewhat long and time consuming.

An alternative to the CIS is a new scale developed by the University of Pittsburgh Office of Child Development called the Caregiver-Child Social-Emotional-Relationship Rating Scale (CCSERRS), which is an amalgamation and a distillation of caregiver/teacher-child interaction dimensions present on the CIS, ITERS and ECERS, HOME Inventory, and others (McCall, Groark, & Fish, 2007). The CCSERRS has the advantages of focusing on major dimensions that underlie most other prominent scales of caregiver/teacher-child interactions, and can be conducted in a few minutes if given as a complement to the ECERS. Its disadvantages are that it is a new scale that has not been tried in early care and education environments (it was developed in orphanages) and it does not include items that reflect the nature of teaching (teaching techniques) although it does reflect the social and emotional style with which the teacher performs these activities. Nevertheless, we recommend using the CCSERRS.

Other indices of quality. A variety of indices of quality of programs can be obtained through questionnaires of the site, including teacher education and experience, group size, teacher:child ratios, use of curriculum, etc. Table 2 presents a list of questions that each site could be asked that would provide these other indices of quality.

Characteristics of children and families. Programs will vary in the nature of the children and families who have selected to attend different program types. It will be useful to ask each parent when obtaining informed consent a few questions about their child and family that can be used to describe the nature of the children and families who participate in each type of program, to adjust the effects on children's development in each type of program for these differences in participants, and to use this information to determine the relative contribution of general family characteristics versus program characteristics on children's development as assessed in this project. Table 2 presents suggested questions for parents in this regard, and these questions have been selected because the literature has shown their relations to early childhood development.

Table 2. Possible Program and Child Characteristics to Be Considered

Program Characteristics

1. Keystone STARS, Head Start, Pre-K Counts, Non-State Supported? Is this unit administered by a school district?
2. For child care programs, family (≤ 6 children), group (7-12 children), or center (13+ children) program?
3. What does the site use the state money for? (i.e., put in general program support, increase number of children, pay for low-income children to attend, increase the length of day children are served, improve quality, increase collaboration)
4. If Keystone STARS, what services received (i.e., professional development, technical assistance) for quality improvement; did unit advance a star in last year?
5. Urban, rural location?
6. PA Region?
7. Caregiver education level, group size, teacher:children ratio, turnover of staff last year, use of curriculum in the classroom/group of children who are assessed.
8. Staff were trained in the last two years using the Mind in the Making Video Curriculum?
9. Predominate income level of families using the site?
10. Total number of teachers for and children of 3-5 years of age at this site.

Child/Family Characteristics

1. Parent education, household earned income? Does parent get government subsidy for child to attend site?
2. How many adults and children live in your household?
3. Was your child birth weight less than 5 lbs. 8 ozs?
4. How old were you when your child was born?
5. Child has disability (i.e., child has Individual Education Plan); obtain specific disability from state records; which special services has child received in last year (i.e., speech/language, physical therapy and support, hearing services, visual

- services, mental development, behavioral management, medical, etc.); from which units were services received (i.e., itinerant services, early childhood unit, early childhood special education, Head Start, Title V services)?
6. Has child attended the current service (Keystone STARS, Head Start, child care) before and for how many months, days per week, hours per day)?
 7. Has the child or family ever attended or received the following services in addition to the current program?
 - a. Nurse-Family Partnership
 - b. Parent-Child Home Program
 - c. Head Start – ages, extent
 - d. Child care (private), ages, extent
 8. Is the main language spoken at home?
 - a. English
 - b. Spanish
 - c. Other
 9. What is the child's predominate racial/ethnic group?
 - a. White
 - b. African-American
 - c. Hispanic
 - d. Asian-Pacific Islander
 - e. Native American
 - f. Other
 10. How many hours per week does child attend this site? Other services?
 11. How many different nonparental care services has your child attended in the last year?
-

Assessment Instrument Recommendation – Children’s Assessment

Generally, we recommend using the **Bracken Basic Concept Scale – Revised** as the assessment instrument for the proposed evaluation. In particular, we recommend using the *School Readiness Composite* set of Subscales (Colors, Letters, Number/Counting, Sizes, Comparisons), supplemented with the *Quantity* and the *Time/Sequence* Subscales.

Rationale and justification. The Bracken meets all of the criteria for the assessment instrument outlined above:

- *Single instrument or subset for all children.* The format of the Bracken simply requires children to point at a particular picture that displays what the examiner requested orally of the child. Thus, while children must be able to hear and comprehend the examiner’s oral questions, they do not need to speak, write, or even point (a child could blink their eyes when the examiner pointed to the correct alternative). So children with a variety of disabilities would not be limited in taking this test. Further, since spoken language is not required of the child, language production and speech limitations are also not a problem (although the test must be administered in a specific language).
- *Independent examiners.* The Bracken is given by independent examiners, not teachers or parents who have limitations for the current purposes that are described above.
- *Balanced content reflecting practical and conceptual abilities.* Table 3 presents a summary of all of the Subscales contained in the entire Bracken series. The first six Subscales (i.e., Colors, Letters, Numbers/Counting, Sizes, Comparisons, Shapes) constitute a *School Readiness Composite (SRC)*, which may be supplemented with one or more of five other Subscales (i.e., Directions, Positions, Self-/Social Awareness, Texture/Material, Quantity, Time/Sequence). The summaries of the Subscales in Table 3 show that the content of these Subscales is clearly the kinds of practical skills and information likely to be taught in most early care and education programs. In contrast, they lack the abstract/conceptual/thinking items on IQ tests that are likely less influenced by early childhood programs. Further, the School Readiness Composite was especially designed to reflect skills needed for kindergarten and primary school, and research indicates that the SRC is related to early school performance. In particular, the SRC accounted for between 45%-65% of the variance in the Metropolitan Readiness Test – 6th Edition with kindergarten students from the fall of the school year to six months later. Also, the SRC has been found to predict academic success for at-risk students, better for black than for white children, and better than the Bracken Total Test Score. Thus, the Bracken appears to be suitable as an assessment of school readiness. See Appendix I for a general review of examples of different types of assessments and Appendix II for summaries of professional reviews of the Bracken.

- *Reflect Pennsylvania's Early Learning Standards.* Table 4 lists each of Pennsylvania's Early Learning Standards categories and subcategories, the number of specific standards listed under each subcategory, and the number of those standards covered by one or more Subscales in the entire Bracken as determined by the Pennsylvania Department of Education and Psychological Corporation of Harcourt, publisher of the Bracken. Note that of the 27 different subcategories of Early Learning Standards, 23 (85%) are represented by at least one Bracken Subscale; at least 30% of the specific individual standards within the eight broader categories of standards are represented by a Bracken Subscale; and over all 136 specific standards, 60 (44%) are represented by at least one Bracken Subscale. Thus, the total Bracken Basic Concept Scale is broadly representative of Pennsylvania's Early Learning Standards. In addition, a review of the child/teaching/assessment strategy emphasized by OCDEL for early care and education programs (i.e., Work Sampling System and the Ounce Scale) emphasizes a variety of early learning skills to be promoted within their curricula, and many are included on the Bracken; further, nearly all of what is represented on the Bracken is included in these curricula.
- *National norms, standardized scores, and percentile ranks.* Scores on the School Readiness Composite, Total Score, and on each Subscale can be converted into a scaled score on a single scoring dimension (i.e., scores on each Subscale, Composite, or Total Score are comparable to one another), as well as percentile ranks in a national standardization sample.
- *Short administration time.* The School Readiness Composite can be administered in 10-15 minutes per child, while the entire set of Subscales might take 30-40 minutes (see discussion below of shortening or expanding the Bracken).
- *Appropriate for children 36-60 months of age.* The Bracken is specifically designed for children 2.5-7.9 years of age, and especially children six years of age and under.
- *Have good psychometric information on reliability, validity, and predictability.* Psychometric information on both the Total Basic Concept Scale and the School Readiness Composite is provided in Appendices I and II. Generally, the test is reliable (i.e., produces the same score for the same child on two different assessments), is internally consistent (i.e., one part of a Subscale produces a score similar to another part of the same Subscale), and is valid in the sense that it correlates (i.e., places different children in the same relative rank ordering) as much larger similar tests, tests of general intelligence, and assessments of school performance.

Alternatives, reductions, supplements. There are several alternatives and supplements to the Bracken and its School Readiness Composite.

- **Woodcock-Johnson III.** One of the most prominent alternatives to the Bracken is the Woodcock-Johnson III test that includes a great variety of Subscales, some of which are similar to those on the Bracken. A subset of Subscales from the Woodcock-Johnson could be an alternative to the Bracken (the entire Woodcock Johnson takes 1-2 hours). An examination of its Subscales, however, indicates that only a few are suitable for preschool-aged children, some require audiotape stimuli, and most scales quickly reach into higher skill levels than we are likely to see in most 3-5 year-olds. The Woodcock-Johnson scales would provide better discrimination and program sensitivity at high skill levels, but less sensitivity at lower skill levels. Since many of the Commonwealth's supported programs are aimed at higher-risk, lower-income children who are likely to begin their early care and education experience at somewhat lower skill levels, the Bracken may be preferred.
- **Supplement Bracken with a picture vocabulary test.** One apparent limitation of the Bracken is that while it has letter recognition, it does not have a word recognition or picture vocabulary Subscale (which the Woodcock-Johnson has in two different forms), the Bracken is sometimes supplemented with a picture vocabulary subscale from another test. But the Bracken itself correlates quite highly with the Peabody Picture Vocabulary test, the most prominent among the various picture vocabulary assessments available. Thus, it appears that perhaps the Bracken does not need to be supplemented in this way.
- **Supplement with subscales from the Woodcock-Johnson.** Another possible supplementation would be to add certain other subscales from the Woodcock-Johnson or to use them as replacements for certain scales on the Bracken. For example:
 - Woodcock-Johnson *Understanding Directions* is similar to the Bracken *Directions/Position* Subscale in asking children to understand locative prepositions ("where is the cow under the tree," "point to the dog beside the barn," etc.) and the Woodcock-Johnson Subscale also tests knowledge of negatives ("where is the owl not in the tree") and sequences ("point to the dog and then to the cat").
 - The Woodcock-Johnson *Applied Problems* Subscale is similar to the Bracken *Quantity* Subscale, but the Woodcock-Johnson requires children to identify one, two, three, etc. objects and to perform elementary addition and subtraction as well as telling time, whereas the Bracken *Quantity* Subscale has no true addition or subtraction but is more general in assessing quantitative concepts beyond counting (i.e., many, full, empty, alone).

Generally, the Woodcock-Johnson is a well-regarded test that has been used in certain national assessments. However, it assesses more advanced skills than the Bracken, and it has fewer items at the basic skill levels that are likely characteristic of the low-income

higher-risk children we are likely to assess. Also, there is some advantage to using Subscales from a single measure. The standardization sample is the same for all Subscales, percentile ranks are provided for combinations of Subscales (such as the School Readiness Composite subset), it is less expensive financially, and it requires less training of assessors. For these reasons, we recommend that the Bracken not be supplemented by the Woodcock-Johnson Subscales, even though the Woodcock-Johnson possesses some advantages, especially assessing and discriminating higher levels of skill.

Supplement the Bracken SRC with other Bracken Subscales. There is much to recommend simply using the Bracken School Readiness Composite of six Subscales presented at the top of Table 3. This composite has been researched as a composite; it correlates well with other tests, school readiness, and school performance measures; it can be given in 10-15 minutes per child (not a trivial practical asset); and it assesses obvious basic skills and concepts that one would expect early care and education programs to promote in young children.

However, there is also good reason to supplement it with other Bracken Subscales, primarily because some of the skills represented by those other subscales are frequently matched with Pennsylvania's Early Learning Standards. For example, Table 5 presents each Bracken Concept Subscale, the number of items on that Subscale, the number of Pennsylvania Early Learning Standards covered by that Subscale (out of a total of 136 specific standards), and the number of standards for which the Subscale is the only Subscale or one of only two Subscales representing that standard. Thus, the right-hand column reflects the extent to which the particular Subscale is minimally duplicated by other Subscales in covering various Early Learning Standards (i.e., the higher this number, the more Early Learning Standards are covered *uniquely* or *nearly uniquely* by the Subscale).

As one can see from Table 5, each of the 11 Bracken Subscales covers at least 19 and more likely 25-30+ Early Learning Standards. Assuming the School Readiness Composite of six Subscales would be retained in its entirety because the Composite is widely used for these purposes and has psychometric information and percentile ranks for the Composite score, the issue is whether it should be supplemented by one or more of the other Subscales. Obviously, each of the other Subscales is related to at least 25 out of 136 early learning standards, and most of the other Subscales can be judged as having items that reflect higher levels of skills than those in the School Readiness Composite, thus providing greater sensitivity and discrimination among higher levels of skills. However, the five additional Subscales have a total of 220 items compared to the 88 for the School Readiness Composite, and while a single child will not be posed all of those items, one does risk extending the length of assessment time to an unacceptable extent if all of the other Subscales were included. In that regard, the right-hand column indicates that the Subscales of Quantity and Time/Sequence are much more likely to be the only or

Table 3. A Summary of the Subscales of the Bracken Basic Concept Scale – Revised

The following examples give some idea of the nature of each subscale. Illustrative items tend to be those in the subscale that are most appropriate for younger children.

School Readiness Composite:

Colors – The child is asked to point to swatches that are a specific color (“Show me which color is black, green, pink, etc.”).

Letters – The child is asked to point to various letters written on a card (“Show me the A, the X, the S, etc.”).

Numbers/Counting – The child is asked to point to one, two, three, or more objects (“Show me the one bear, nine bumblebees, three flowers, etc.”).

Sizes – The child is asked to identify one object that is of a different size than another object in a picture (“Show me which animal is big, which ball is little, which dog is small, etc.”).

Comparisons – The child is asked to identify pictures of objects that are different by comparison with other objects (“Show me which boxes are not the same, which fruit are different, which puzzle pieces fit exactly, which shoes match, which boats are alike, etc.”).

Shapes – The child is asked to identify pictures of certain shapes (“Show me the star, the heart, which children are in a line, the circle, the cone, etc.”).

Other Subscales:

Directions/Positions – The child is asked to identify a picture specified by a certain position, locative preposition, or direction (“Show me which boy has his hat off, which child is on the swing, which door is closed, which clown is upside down, which chicken is inside the house, etc.”).

Self-/Social Awareness – The child is asked to identify pictures in which people are displaying certain emotions, represent certain states, or have certain social identities (“Show me which child is crying, which child is sick, the girl, which puppy is resting, which child has been hurt, which person is relaxing, etc.”).

Texture/Material – The child is asked to identify pictures of objects that have certain textures or physical characteristics (“Show me which one is heavy, which child is making a loud noise, which one is hot, which one is made of wood, which one is sharp, which one is boiling, etc.”).

Quantity – The child is asked to identify an object among alternatives by virtue of its quantitative characteristics (“Show me which tree has many apples, the dollar, which box is empty, where both dogs are asleep, which bird has nothing to eat, a whole pie, etc.”).

Time/Sequence – The child is asked to identify an object or scene among alternatives that is characterized by its time or sequence (“Show me where it is night, which child is fast, which child has finished drinking, which shoes are new, which person is leaving the store, where a cupcake has been skipped, where it is morning, etc.”).

**Table 4. Pennsylvania Early Learning Standards
Covered by the Bracken Basic Concept Scale – Revised**

<u>Standard Category/Subcategory</u>	<u>Number of Standards</u>	<u>Standards Covered by Bracken</u>	
<u>Approaches to Learning</u>			
1. Initiative and curiosity	4	1	
2. Engagement, persistence	3	2	
3. Flexibility, risk taking, responsibility	5	2	
4. Learning, problem solving	3	1	
5. Imagination, creativity, invention	<u>2</u>	<u>0</u>	
	17	6	35%
<u>Language and Literacy</u>			
1. Listening, understanding	4	3	75%
<u>Expressive Language</u>			
1. Communicates ideas, experiences, feelings	8	3	38%
<u>Comprehension</u>			
1. From written, oral stories	7	6	86%
<u>Literacy</u>			
1. Sounds of language	8	2	
2. Awareness of concepts of print	9	5	
3. Book knowledge and appreciation	5	1	
4. Understands letter knowledge	3	2	
5. Drawing letter-like forms, etc.	5	0	
6. Writing as communication	<u>3</u>	<u>0</u>	
	33	10	30%
<u>Logical Mathematics</u>			
1. Numbers, numerical operations	7	6	
2. Patterns, relations, functions	6	4	
3. Space and shape	6	4	
4. Measurement concepts	3	3	
5. Represents and interprets data	4	1	
6. Reasons, predicts, problem solves	<u>6</u>	<u>2</u>	
	32	20	63%
<u>Science</u>			
1. Process of scientific inquiry	7	2	
2. Characteristics of living things	4	2	
3. Physical properties of objects	6	2	
4. <u>Earth and space</u>	<u>6</u>	<u>1</u>	
	23	7	30%
<u>Social Studies</u>			
1. Self within community	8	2	

2. Past, present, future	3	3	
3. Role of consumers	<u>1</u>	<u>0</u>	
	12	5	42%
Total	136	60	44%

Table 5. Bracken Concept Subscale and Frequency of State Pre-K Standards

Bracken Subscales	Number of Items on Subscale	Number of Standards Covered out of 136	# of Standards for which Subscale is only 1 or 2
<u>School Readiness Composite:</u>			
Colors	11	21	0
Letters	16	19	3
Numbers/Counting	19	25	3
Sizes	12	32	1
Comparisons	10	33	3
Shapes	<u>20</u>	28	1
Subtotal =	88		
<u>Other Subscales:</u>			
Directions/Positions	65	25	1
Self-/Social Awareness	38	25	3
Texture/Material	31	32	5
Quantity	49	38	10
Time/Sequence	<u>37</u>	37	12
Subtotal =	220		
Total	308		

nearly the only assessment of a Learning Standard than any of the other Subscales, so if the School Readiness Composite is to be supplemented, it would appear that the Quantity and Time/Sequence Subscales would represent the best choice in terms of providing more comprehensive coverage of the state’s Early Learning Standards.

Supplement with a social-emotional-relationship assessment. A major Early Learning Standard and component of the child/teaching/assessment strategies emphasized by OCDEL for early care and education programs in the state focuses on the social-emotional-relationship development of young children, but the recommended Subscales do not include an assessment of this domain. The Bracken has a Self-/Social Awareness Subscale, but this is largely an assessment of the child’s ability to identify various

emotions (crying, sad, hurt) rather than children's social-emotional interactions and self-regulation which are more directly related to school readiness and mature social relationships. Indeed, a limitation of independent-examiner administered assessments is their relative inability to assess social-emotional-relationship development (at least without asking the teacher, which has certain limitations in the current context as described above). Therefore, we suggest that attempts to assess this domain have lower priority, because it cannot be accurately or comprehensively accomplished within the current circumstances.

Assessment Instrument Recommendation – Classroom Quality

General environmental quality. We recommend that the ECERS and FDCERS be used because OCDEL already requires these assessments for certain early care and education programs in the Keystone STARS system. This means that OCDEL has decided that they represent as well as any instrument the criteria for quality that the Commonwealth desires in early care and education environments. Also, it is possible to compare the project's ECEERS/FDCERS results with data collected before the state started to support any early care programs.

The project could accept the scores of OCDEL administered ECERS and FDCERS that may have been conducted on the teachers in classrooms that had been selected from Keystone STAR sites if such assessments were done by OCDEL personnel (not by the early care site itself) in March, April, and May and the summer before pre-testing or during the pre-test – post-test interval. However, the number of such assessments is expected to be only a small proportion of the total number of sites, the OCDEL assessments would be conducted by different assessors, they might deviate in time from the others, they would only be available for certain keystone STARS programs (a confound), and OCDEL would not have given the CCSERRS teacher-child interaction scale (see below). Therefore, we recommend the project conduct all ECERS and FDCERS.

Teacher-child social-emotional-relationship characteristics. We recommend that the Caregiver-Child Social/Emotional/Relationship Rating Scale (CCSERRS) be conducted as a supplement to the ECERS and FDCERS to complement the ECERS/FDCERS with an assessment of the teacher-child interactions. While the CCSERRS is untested in early care environments, its creation was based on all of the prominent assessments of teacher or caregiver-child interactions and represents the major dimensions of these different instruments. Further, once the ECERS or FDCERS has been administered, it takes only a few minutes for the assessor to fill out the CCSERRS ratings of 18 caregiver and child characteristics. While the CCSERRS does not reflect teaching techniques, it is a direct rating of the social and emotional interactions between teachers and children as well as the teacher-child relationship in the context of the early care setting. See Appendix III for a description of the 18 CCSERRS ratings.

Procedure and Timeline

One or Two Waves of Data Collection?

The minimum sample described in Table 1 could potentially be accomplished with one Wave of data collection beginning in late August for 4-5 weeks and finishing in the following May. This would require three assessment headquarters, one in the east, center, and west of the state, and each assessment headquarters would need two coordinators to contact sites, set up parental informed consent, select children, arrange assessment appointments, and make travel arrangements for 8 assessors at each headquarters. Each assessor would need to visit approximately 5-7 sites per week and assess 20-25 children per week to obtain the minimum sample in Table 1 after attrition. However, this approach produces only a minimum sample and one risks not obtaining the minimum sample because of lack of site cooperation, parental/child absences, inability to obtain informed consent, lack of children in selected sites that meet the eligibility criteria and sampling preferences, etc. Since all pre-testing of nearly 2,000 children in approximately 550 sites would need to be accomplished in approximately 5 weeks of pre-testing, conducted by 24 assessors, a single Wave of data collection would be a high-pressure activity with little tolerance for the procedural irregularities that inevitably occur in such field work.

Two Waves of data collection—in which Wave I occurs in the fall and spring of one year and Wave II occurs in the fall and spring of the next year, would have several advantages:

- It would be able to compensate for the consequences of sampling and procedural irregularities experienced in Wave I with another year of sampling and assessments in Wave II.
- The sample size could be increased by 50% or more which would allow statistical analyses to compare children within program type groups on the various program characteristics listed in Table 1 as well as individual differences in both program characteristics and children within program types.
- Certain comparisons that are unlikely to be possible in a single data collection Wave would be possible in a two Wave study, including comparing full-day with part-day attending children and directly assessing children who spend more than 7-8 months in the same program who are more likely to show program developmental benefits.

Of course, a two data collection Wave study is somewhat more expensive (see budget estimates below) and takes longer.

Procedures and Timelines

Table 6 presents a possible set of procedures and a timeline for the proposed project. Items in Roman type in Table 6 are for a project with a single data collection Wave; items in italics would be added if the project involved two data collection Waves. The dates in Table 6 are predicated on a project that would begin January 1, 2008 ending February, 2010 for a single Wave study or January 31, 2011 for a two Wave project.

Each Wave of data collection involves conducting pre-tests on children in approximately September, conducting quality assessments of classrooms in which children were given pre-tests between October and April, and then conducting post-tests on children in approximately May. Thus, there would be 7-8 months of program exposure between pre-tests and post-tests. If a two Wave data collection project were conducted, Wave I children who are still in the same program in the second year of data collection will be given a Follow-Up Assessment in February-April 2010. These children would have at least 12-13 months of exposure to the program up to a maximum of 18-20 months of exposure to the program depending on whether they remained in the program through the summer and how late in 2010 they were assessed. This provides a direct test of whether children exposed longer to the program derived more benefit from it than those who are exposed a shorter period of time. It is also possible that program benefits may not be readily detected in children with only 7-8 months exposure but be more substantial and detectable in children exposed for more than one academic year. Figure 1 presents a simplified possible timeline for a one- and a two-Wave study.

Data collection procedure. Sites should be contacted by telephone to invite their participation and to obtain basic program characteristics (e.g., see Table 2). Arrangements for a place to assess children must be established. When it is clear which children will be enrolled in approximately August-October, all children who meet the project's sampling preferences (see sampling steps 3 and 4 above) will be identified for assessments. Plans to obtain parental permission and family information will then be made, preferably by site staff. Assessments will then be scheduled for a single or consecutive days. Procedures will be similar for post-testing and follow-up testing.

Data Analyses

Classroom Quality Assessments

Each sampling unit in Table 1 would have an assessment of classroom or group quality, the ECERS or FDCERS plus the CCSERRS of teacher-child social/emotional/relationship interactions. The ECERS/FDCERS yield subtest scores which can be analyzed as a multivariate set and a total score that can be analyzed as a single variate. The CCSERRS yields a total score that can be analyzed as a single variable, and the scale could be factor analyzed to determine underlying themes in this population which could then be used to create subscale scores (the CCSERRS has been factored before, but for orphanage caregivers which may yield a different factor structure than these early care teachers).

Table 6. Possible Project Procedures and Timeline
(Items in italics are for a two-wave data collection project)

January – August 2008

1. Project starts
2. Measurements selected and questionnaires created
3. Institutional Review Board documents prepared, submitted, approved
4. Sampling is determined, Wave I sites are contacted, and characteristics determined
5. Assessors hired and trained
6. Assignments of assessors to sites determined, travel arrangements planned
7. Database developed, machine readable scoring forms created, data input program written and tested
8. Meet with OCDEL administrators to finalize plan

August – October 2008

1. Wave I Pretests are administered to children in all Wave I sites

October 2000 – April 2009

1. Data are entered into database and cleaned; missing data handled
2. Descriptive data analyses conducted on Wave I Pre-test data
3. Post-testing schedules created
4. Classroom quality assessments conducted on all classrooms having children in Wave I Pre-test.

April – June 2009

1. Wave I Post-tests are administered in all Wave I sites

June – August 2009

1. Wave I Post-Test data are entered into database and cleaned; missing data handled
2. *Preliminary analysis of Wave I data conducted to determine sampling inadequacies to be remedied in Wave II.*
3. *Wave II sites contacted and characteristics determined*
4. *Wave II Pre-Testing schedules created and travel arrangements made*

August – October 2009

1. Wave II Pre-Tests are administered at all Wave II sites.

October – December 2009

1. Wave I Pre- and Post-Test data are analyzed
2. Brief preliminary report on Wave I data is written; meet with OCDEL administrators to discuss preliminary results
3. *Wave II Pre-Test data are entered into database and cleaned; missing data dealt with*
4. *Classroom quality assessments conducted on all classrooms having children in wave II Pre-Test*

January – February 2010

1. Wave I data analysis completed, meet with OCDEL to discuss results, write report.
2. *Wave I Follow-up and Wave II Post-Test scheduling is planned*

February – April 2010

1. *Wave I Follow-up assessments administered on all children in all Wave I sites who had a Wave I Pre-Test (Fall 2009) and who are still in the same program and site in March-April (2010)*

April – June 2010

1. *Wave II Post-Tests are administered to all children in all Wave II sites who were given Wave II Pre-Tests in August-October 2009*
2. *Wave I Follow-up data are entered into database and cleaned; missing data dealt with*

June – October 2010

1. *Wave II Post-Test data are entered into database and cleaned; missing data dealt with*
2. *Data analyses conducted*
3. *Meet with OCDEL administrators to discuss results*

November – December 2010

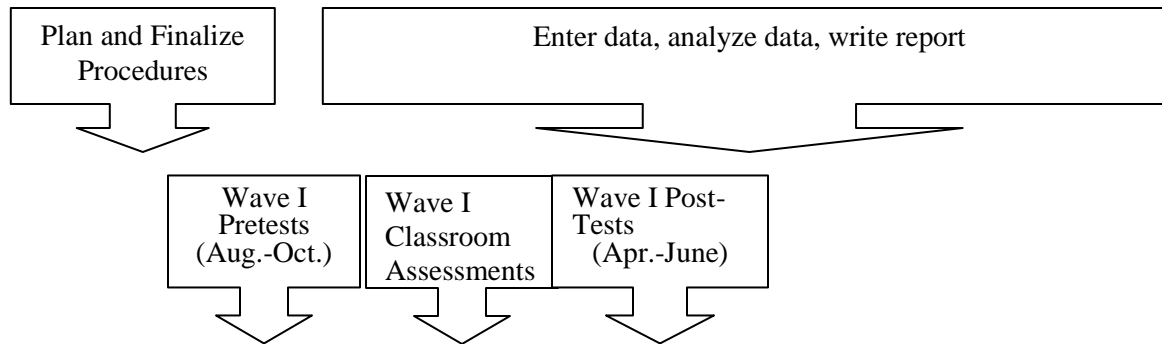
1. *Write Final Report*
2. *Meet with OCDEL administrators to discuss draft of Final Report*

January 2011

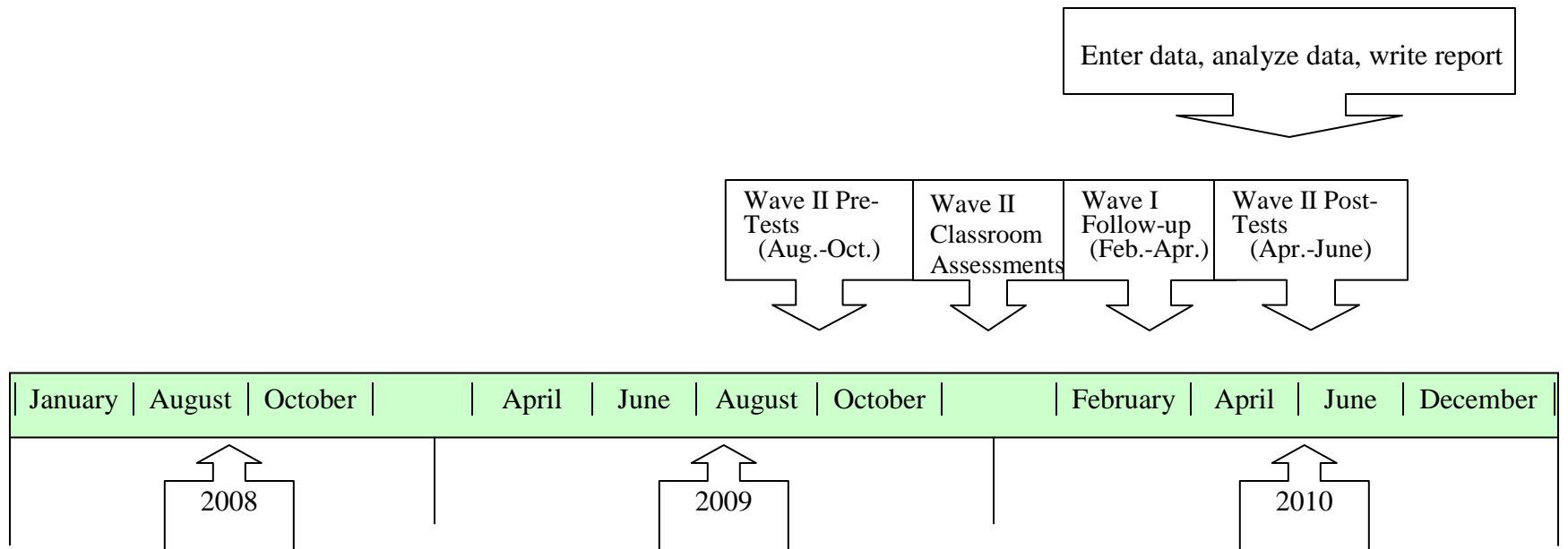
1. *Release Final Report*
-

Figure 1. Simplified Possible Timeline for One- and Two-Wave Studies

Wave I:



Plus Wave II:



The basic analysis is to compare the average classroom quality scores, whether a single variable or a set of subscales, between each of the sampling units in Table 1. For example, one would expect Keystone STAR centers to have generally higher quality scores than non-state supported child care, and for classroom quality to increase with STAR level (because the ECERS/FDCERS are part of the criteria for awarding STAR level—do they also increase in the quality of social/emotional/relationship interactions between teachers and children?). Such results can be compared with OCDEL’s assessment of STAR levels using the ECERS in 2006. Similar analyses can be conducted for each of the major program types versus the non-supported child care programs, and different types of programs can be compared with each other. Such analyses will reflect classroom quality differences between program types (e.g., does Head Start provide a generally higher quality classroom for low-income children than Pre-K Counts, and is the quality of Head Start as good as STAR 3 or STAR 4 programs that likely serve children from somewhat higher income families. It will also be interesting to observe the correlation between ECERS/FDCERS scores and CCSERRS scores—do sites that have better environmental quality also provide warmer, more caring, sensitive, and responsive teacher-child interactions? It will also be useful to observe whether classroom quality varies as a function of the income level of the families of children who attend (i.e., do the advantaged children attending state-supported programs get a better quality of care than the disadvantaged children, or are state-supported programs leveling the early childhood playing field for children from different economic circumstances?).

Classroom quality on the ECERS/FDCERS can also be related to similar scores for center, group, and family child care Head Start programs and licensed nursery schools conducted in 2002 before Pennsylvania began to support programs. Has quality improved since 2002 and with state support?

The classroom quality assessments can also be used in analyses of the children’s developmental data (see next section). For example, once classroom quality is considered, do children develop differently in one versus another program or context? It is possible that once the child characteristics and program quality are taken into account, children develop similarly regardless of which program they are in. If that were the case, it would tell the Commonwealth to improve program quality as best it can in all program types.

Children’s Assessments

Percentile ranks. The children’s assessment instrument and its norms will provide standardized scores and **percentile ranks**, the latter being *the percentage of children in the national standardization sample of the age at assessment of the target child who would be expected to score lower than the target child*. A percentile rank of .88 means that 88% of children of that age would be expected to score lower than the target child (i.e., higher percentile ranks reflect better performance). Percentile ranks are age invariant, which means that pre-test scores can be compared with post-test scores for children in different programs and at different ages. Average percentile ranks of .50 would mean that children in such a program were approximately “average” (technically

at the median) relative to children across the country who might be enrolled in a variety of different programs or not enrolled in any program. Each child will have a pre-test and a post-test percentile score, and a child who obtains the same percentile rank at post-test as he or she did at pre-test would be making typical developmental progress. A post-test percentile rank higher than the child's pre-test rank indicates the child developed more rapidly than most children during the interval between assessments, whereas a child whose post-test percentile rank is lower than the pre-test ranking would be progressing more slowly than the typical child.

Program effect scores. It is also possible to calculate a **program effect** score, which basically reflects *the amount of developmental progress children make relative to the amount of progress we would expect specifically of children in that particular program type or with the particular characteristics (i.e., low-income) of interest if they did not experience the program.* Several program effect scores can be calculated for an individual child depending on which group(s) the child belongs to. That is, program effects could be calculated separately for each program type, for each income level of children, etc. Program effect scores permit comparisons to be made specifically to Pennsylvania children who enroll in these particular programs or who have certain characteristics (i.e., low income), who may be different in important ways from the very comprehensive and general standardization sample that provides percentile ranks.

A program effect score (McCall, Ryan, & Green, 1999) is calculated in the following way. The pre-test scores for all children in a specific group (all children in the entire study, all Head Start children, all children with family incomes in the poverty range, etc.) are plotted as a function of the age of the children at the pre-test assessment. Assuming children enter programs at different ages between 36 and approximately 48+ mos., these scores describe the relation between age and percentile ranks (or raw or standard scores) for that group of children who had not yet experienced the particular early care and education program. If a national sample of children were obtained, the relation between percentile rank and age would be a horizontal line at the 50th percentile rank. But such a relation is not necessarily an accurate portrayal of the performance of children in Pennsylvania programs or each subgroup of the population. That is, children in Head Start may have a generally lower percentile rank than children in Keystone STARS programs having 3 and 4 STARS. Also, the literature on low-income children indicates that such children tend to decline in percentile rank over the first 5-6 yrs. of life (e.g., McCall, 1993), so for low-income children we would actually expect them to have a lower percentile rank on their post-test than on their pre-test if they were not exposed to an early care and education program.

Then an expected no-program score is calculated by predicting for each individual child (using their pre-test score and the relation between all pre-test scores and age) that child's post-test score given the child's months of exposure to the program. Then the child's actual post-test score minus his or her predicted no-program score equals the program effect. Children with a positive program effect score improved while they were in the program relative to what one might expect of children in that child's group if they had not been exposed to the early care and education program. The program effect score

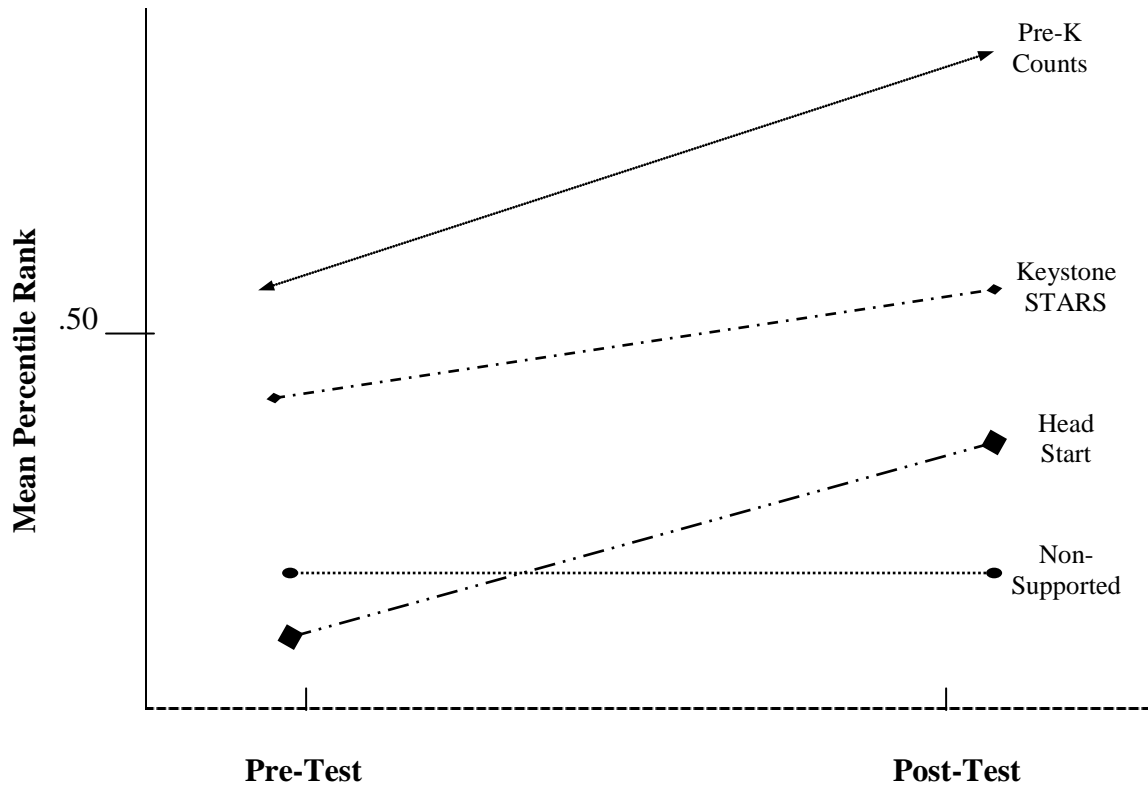
strategy can help to equate different groups of children who, without exposure to early care and education, might be expected to develop at different rates, and it helps equate different state-supported programs that otherwise differ in the backgrounds and skills of children who tend to be enrolled in one versus another program (see examples below).

Basic pre/post analyses on percentile ranks. The simplest presentation of study results is to plot and statistically analyze the mean pre-test and mean post-test percentile ranks for children in each type of program or program characteristic. For example, all children in Head Start, Keystone STARS, Pre-K Counts, and non-supported programs might have their average pre- and post-test percentile ranks plotted on a graph similar to that presented in Figure 2. These hypothetical results (which may not turn out to be the case at all) at least illustrate what such a simple pre- versus post-test percentile rank analysis might show. The graph suggests that children in non-supported programs and in Head Start have generally lower percentile ranks than children in Keystone STARS; children in Pre-K Counts have the highest average percentiles. Children in the non-supported programs do not change in their average percentile rank from pre- to post-testing, whereas children in Head Start, Keystone STARS, and Pre-K Counts do develop at a faster than typical rate relative to the standardization sample (presumably the US population of children exposed to all kinds of programs and no programs). Such plots can be made separately for other program types, such as, sites of different STAR levels in Keystone STARS; family, group, and center children in child care programs; or for children from families of different education and income levels.

Of course, we know that within each of these major program types, program and child characteristics may influence the extent to which a child improves or declines from pre-test to post-test. Statistically, the child's post-test score is used as the outcome measure and the child's pre-test score is used as a statistical covariate, which means that the statistical analysis essentially equates all children on their pre-test scores and asks whether the post-test score is higher or lower than might be expected on the basis of the child's pre-test score. The child's adjusted post-test score is then related to program characteristics and individual child characteristics, and the analysis demonstrates how much change in post-test score is associated with program type, program characteristics, classroom quality, and child characteristics (see Table 2). The unit of analysis is the individual child coupled with standard error adjustments for the fact that some children are in the same classrooms—then the analysis is basically a multiple regression with planned sequence of entering predictors.

Basic pre/post analyses on program effect scores. Similar pre- and post-test plots can be made using the program effect score rather than the child's percentile rank. The reason one might do such analyses is because children who typically enroll in one type of state-supported program may be different both in their pre-test developmental level and in the developmental progress that would be expected without residence in the program. As stated above, for example, high-risk, low-income children are known to decline in percentile rank during the pre-school years, so it is possible that children who enter Head Start programs might be expected to decline in percentile rank if they did not experience the Head Start program. If simple percentiles were plotted as above in Figure 2, Head

Figure 2. Hypothetical results of simple pre- vs. post-test percentile rank comparisons between program types.



Start children might show no change from pre-test to post-test, and we would conclude that Head Start did not improve children developmentally. But, if we would otherwise expect such children to decline in percentile rank if they were not exposed to Head Start, then such a result would actually show positive benefits for Head Start because it prevented the decline such children would be expected to show if they did not experience Head Start. The program effect scores, which take into account the expected developmental progression for children in a specific group (not just the standardization sample) who did not experience the program, would reveal this positive benefit for Head Start.

Thus, if program effect scores were calculated separately for children in each program type group, those results might actually be quite different than the simple pre/post percentile rank outcomes because children with different expected developmental performance might enroll in different programs. Also, the program effect analyses may show more substantial increases for low-income children within program types than would the simple pre- versus post-test percentile rank plots illustrated above.

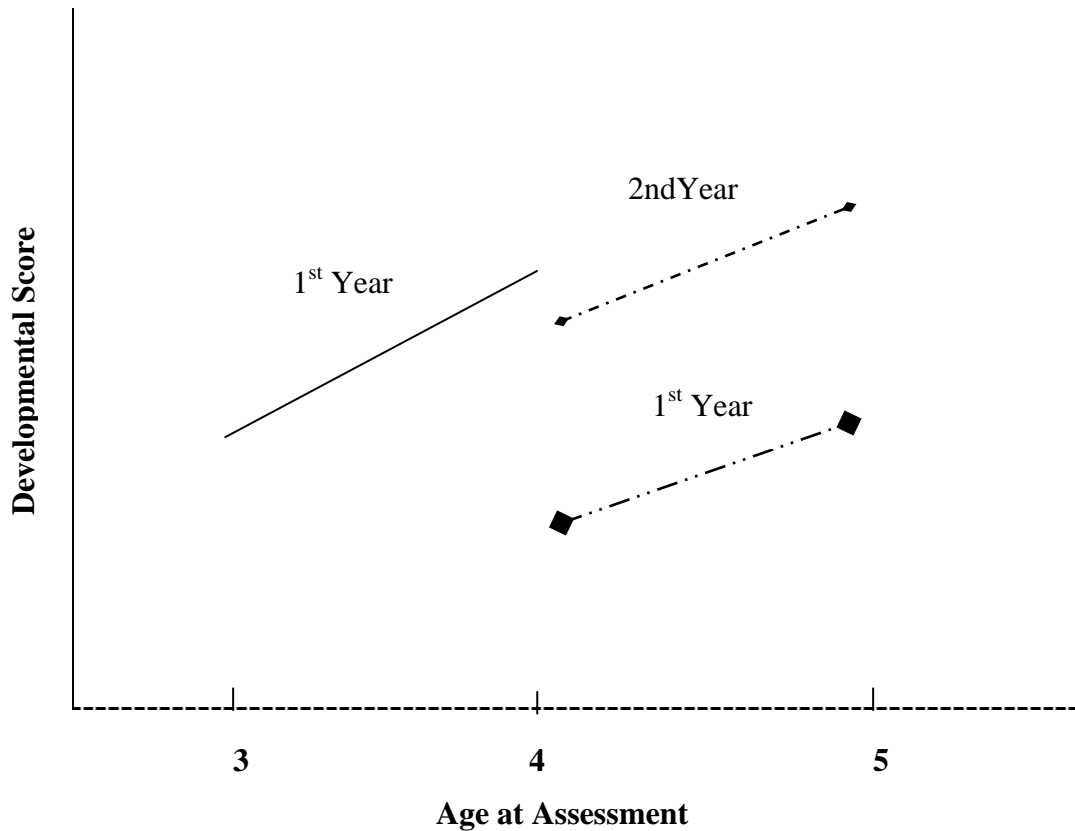
Data analyses would be similar to those described above and reveal the relative contributions of different types of programs, characteristics of programs, and characteristics of children within the specific groups for which program effect scores were calculated.

These several analyses would answer all of the major questions posed at the beginning of this report.

The effect of longer residences in the program. The basic design tests whether children exposed to only 7-8 months of a program type (i.e., 9-month school year minus a month of pre-testing and post-testing) show developmental improvements. What if this is not enough exposure to a program to produce benefits? There are at least two ways to examine whether longer exposures produce more and continuing benefit for children:

- **Wave I follow-up in Year 2.** The most direct method is to assess those Wave I children who are still in the same program and site during Year 2 of data collection in a two-Wave project. Such children could be given a second post-test in early spring of the second study year to determine if they improved between their first-year post-test and their second-year post-test and whether they improve more on their second-year post-test relative to their pre-test after 14-20 months exposure than other children do after 6-7 months of exposure. This approach assumes there are sufficient numbers of three-year-olds in Wave I who remain in the same program/site, and there may be more children in some programs (i.e., Keystone STARS, non-supported child care) than others (e.g., Head Start, Pre-K Counts).
- **Second- vs. First-Year children.** Ignoring the Wave I follow-up described above, all children will get a pre-test and a post-test, but some will be First-Year children (i.e., three or four-year-olds who have had no previous experience with an early learning program (three-year-old are defined to be inexperienced even if they attended the same site the previous year)) and some four-year-olds who are Second-Year children (i.e., they experienced the same program as three-year-olds the previous year). A comparison between these groups may provide evidence of the benefits of longer exposures. Figure 3 shows hypothetical results for First-Year children who begin at 3 or 4 years old and Second-year children who begin at 4 years old. If longer program exposure produces more developmental gain, one might expect this pattern of results in which the Second-Year children finish higher than either of these groups in the same program. This would be the strategy if a single data collection Wave study was conducted.

Figure 3. First-Year children who start the program at 3 or 4 years of age compared to Second-Year children who are assessed at 4 and 5 years of age.



Estimated Budget

One Wave, Total Minimum Sample Budget

The estimated budget for a one-wave data collection project using the minimum sample for all proposed program types (Table 1) is given in Table 7.

Budget periods. The budget is divided into three budget periods of one year each, but the project would only cover nine months of the first budget year, 12 months of the second budget year, and six months of the third budget year. This was done to have budget years that correspond to those prescribed by the Federal Administration for Children and Families (ACF) requirements as laid out in their Request for Proposals (0031) in which the annual budgets run from the last day of September in each year. ACF can provide a maximum of \$250,000 the first year and \$200,000 in each of the two succeeding years (subject to certain constraints). Casting the budget into these time periods permits OCDEL to readily see how much additional money would be required to

supplement the maximum ACF potential contribution (a cost-share of at least 20% of *total project cost—not total federal contribution*) is required by ACF in any case, and any indirect costs must come out of the maximum potential contribution (not in addition to it).

Table 7. Budget for One-Wave Data Collection, 27 Months, Total Minimum Sample

Function	Year 1 (9 Months)		Year 2 (12 Months)		Year 3 (6 Months)		Total
	% Effort	Mos.	% Effort	Mos.	% Effort	Mos.	
Total Project Personnel:							
Principal Investigators	50%	9	60%	12	90%	6	
Project Coordinator, Data Collection Director (Hdqtr. 1)	50%	9	50%	12	50%	3	
Database Mgr., Statistician	50%	9	50%	12	50%	6	
Statistical Specialist			20%	12	40%	6	
Data Entry							
Fringe Benefits							
<i>Subtotal – Total Project Personnel</i>		\$90,739		\$168,901		\$127,113	\$386,753
Data Collection:							
<u>Data Collection Managers (Hdqtr. 2 & 3):</u>	50%	3	50%	9			
	50%	3	50%	9			
Fringe Benefits							
<i>Subtotal Data Collection Managers</i>		\$19,947		\$61,336			\$81,283
Training:							
3 Assessment Trainers 5 days each @ \$750		\$11,250					\$11,250
24 Assessors Training 5 days each @ \$250/day		\$28,800					\$28,800
6 Coordinators Training 3 days @ \$240/day		\$4,320					\$4,320
Travel for Training:							
5 days for 14 Assessors							
3 days for 4 Coordinators							
Hotel @ \$80/day		\$6,560					\$6,560
Per Diem @ \$60		\$4,920					\$4,920
Mileage 16 @ 500 mileage= 8000 miles @ \$0.485		\$3,880					\$3,880
<i>Subtotal Training</i>		\$59,730					\$59,730
Pretest:							
6 Coordinators 45 days @ \$240		\$64,800					\$64,800
24 Assessors 25 days @ \$240		\$144,000					\$144,000
Travel: 24 Assessors @ 25 days, 6000 miles @ \$0.485		\$29,100					\$29,100
24 Assessors 12 days hotel @ \$80		\$23,040					\$23,040
Per diem @ \$60		\$17,280					\$17,280
<i>Subtotal Pretest</i>		\$278,220					\$278,220

<u>Classroom Assessment:</u>				
6 Coordinators 25 days @ \$240		\$36,000		\$36,000
24 Assessors 20 days @ \$240		\$115,200		\$115,200
Travel: 24 Assessors avg. 100 miles/day for 20 days= 48,000 miles @ \$0.485		\$23,280		\$23,280
24 Assessors 10 days hotel @ \$80		\$19,200		\$19,200
Per diem @ \$60		\$14,400		\$14,400
<i>Subtotal Classroom Assessment</i>		<u>\$208,080</u>		<u>\$208,080</u>
<u>Post Test:</u>				
6 Coordinators 35 days @ \$240		\$50,400		\$50,400
24 Assessors 25 days @ \$240		\$144,000		\$144,000
Travel: 24 Assessors 25 days @ avg. 100 miles/day= 60,000 miles @ \$0.485		\$29,100		\$29,100
24 Assessors 12 days hotel @ \$80		\$23,040		\$23,040
Per diem @ \$60		\$17,280		\$17,280
<i>Subtotal Post Test</i>		<u>\$263,820</u>		<u>\$263,820</u>
<u>Tests:</u>				
27 Bracken Kits @ \$375	\$10,125			\$10,125
3600 Score Sheets	\$11,200			\$11,200
27 ECERS/FDCERS @ \$18	\$486			\$486
<i>Subtotal Tests</i>	<u>\$21,811</u>			<u>\$21,811</u>
<u>Equipment:</u>				
3 Computers @ \$1,000	\$3,000			\$3,000
24 Cell phones and service @ \$33.99/mo. Year 1 for 3 mos., Year 2 for 9 mos.	\$2,447	\$7,342		\$9,789
<i>Subtotal Equipment</i>	<u>\$5,447</u>	<u>\$7,342</u>		<u>\$12,789</u>
<u>Supplies:</u>				
Books, learning materials for site as honoraria (540 sites @ \$35)	\$18,900	\$18,900		\$37,800
<u>Consultants</u>	\$6,000	\$6,000	\$8,000	\$20,000
<u>Travel to Washington Meeting*</u>	\$750	\$750		\$1,500
GRAND TOTAL	\$501,544	\$735,129	\$135,113	\$1,371,786

*Required by ACF only

The first budget year pays for a total of nine months of project activity, a maximum of eight months prior to actual conduct of pretesting which would occur approximately in September (actual pretesting is likely to extend longer than four weeks and spill over into the second budget year, but the cost has been placed all in the first budget year with a

proviso of carryover into the second to cover extended pretest assessments). While it may seem that a great deal of planning has already been conducted and that eight months of additional preparation prior to data collection may seem excessive, much additional work needs to be done and is detailed below primarily under Total Budget Personnel and Training of Assessors.

The second budget year would cover conducting the classroom assessments and the children's post-test assessments (again with extended post-testing to be covered with a second-year budget carryover to the third year).

The third budget year actually extends a maximum of six months, during which time any additional post-test assessments on children would be conducted, the data would be analyzed, and a report written and submitted. Six months for these activities, especially with a database this complicated, is a short period of time relative to most projects this large.

Total project personnel. The project is conceived to be operated out of a Project Central location where major project personnel responsible for organizing, operating, managing the project; maintaining the database; analyzing the data; and writing the project report are housed. In addition, there would be three Data Collection Headquarters located in the eastern, central, and western part of the state (one of which is likely to be Project Central) and would be responsible for conducting all of the assessments across the Commonwealth (see Data Collection below).

Principal Investigator(s). The project should have one or more principal investigators who have knowledge of the research literature, wide experience in conducting evaluation projects, knowledge of assessments, expertise in data analysis, experience in relating research to the needs of policy makers, and substantial experience in writing the results of such studies in language suitable for policy makers. During budget year 1, they would be responsible for writing the application to the Internal Review Board governing confidentiality and ethical principles, which is a substantial and new document (not simply this report or even a modification of this report designed as a grant application) and the approval process can easily take several months. In addition, the principal investigators must work with OCDEL (and perhaps ACF) to insure the project meets their needs as it faces the inevitable changes that must be made when the design is actually implemented. In addition, they must provide direction, supervision, and support to all of the other staff. Their percent time and role increases in the second year and they are primarily responsible for overseeing the data analyses and writing the report in the second and third years.

Project Coordinator. This person is charged with coordinating the data collection headquarters and all of the data collection activities. This person would also likely be a *Data Collection Manager* (see below) for the Data Collection Headquarters housed at Project Central, so this individual actually has two roles. They must coordinate the hiring of the assessors, make arrangements for hiring trainers and conducting the training, purchasing tests and equipment, working with OCDEL to obtain lists of providers of all

program types, organizing the sampling, and preparing lists of sampled sites and replacement sites for data collection personnel – all in the seven months preceding training and data collection. Once data collection begins, this person will need to coordinate the Data Collection Headquarters, deal with the inevitable irregularities and problems that arise, help to coordinate obtaining informed consent and dealing with issues of child sampling in addition to coordinating one of the Data Collection Headquarters.

Database Manager, Statistician. This person must help design data collection forms, create and implement an appropriate data input system (which could consist of appropriate data forms and instructions for hourly workers to enter data by hand or develop data forms that can be scanned into the database and then checked because scanning devices are error prone), design a database to receive these data that is compatible with the way data comes from the assessors and how it has to be filed so that it is convenient for statistical analysis, and all of this needs to be done prior to the start of data collection. Once data collection begins, the database manager must ensure that data are entered correctly and missing data are dealt with, and then conduct preliminary analyses as the project progresses to determine whether sampling is deficient in certain program type categories and to calculate the regressions of pretest scores and age as needed for the calculation of treatment effect (see above).

Statistical Specialist. This database is extremely complicated, and while the statistical analyses described above seem straightforward, they are not. Whenever one has as many secondary variables, such as characteristics of programs and children within program types as well as the overlap in programs (i.e., a given site can be both a Head Start and a Pre-K Counts site, etc.) and where child characteristics are unevenly distributed across program types, one has a very substantial data analysis task. Consequently, it is recommended the project involve a statistical specialist in hierarchical linear regression and modeling who not only has experience in analyzing large and complicated databases of this kind (especially ones involving longitudinal data) but also has some knowledge of and experience with the issues in early care and education and policy to guide formulating relevant questions and answering them with appropriate statistical analyses.

Data Entry. Data need to be entered as they are acquired from the pretest, classroom assessments, and post-tests, and this needs to be done as rapidly as possible, especially considering only six months has been allocated in the third budget year for data analysis and report writing. This can be done either by hand by hourly workers or by using computer scanning devices (but these nevertheless require data to be checked because such devices are known to be error prone). Which system is used will be decided by the Database Manager and Principal Investigators, but in either case there will be an expense in the second and third year for this activity.

Percent-Time Estimates. The budget presented in Table 7 includes percent-times for these personnel, but the challenge in staffing the project is that these individuals must work intensely during certain phases of the project and less intensely at other times, and they must be able to adjust their own work schedules to accommodate these variations.

Thus, the percent-time estimates are averages over the number of months in each budget period.

Data collection. Data collection consists of four phases – training of assessors, children’s pretest assessments, classroom quality assessments, and children’s post-test assessments. While the costs for the total project personnel (above) do not vary with the number of sites and children sampled, data collection costs would change if certain program types were omitted or the sample size changed from the minimum sample in Table 1.

Data Collection Managers. These individuals will supervise the coordinators and assessors in each of the three Data Collection Headquarters. One Data Collection Manager will also be the Project Coordinator whose salary is covered above under Total Project Personnel, so the Data Collection Managers for Headquarters 2 and 3 are listed here. These individuals should have experience in organizing field work of this type, which is quite complicated and riddled with irregularities and unexpected circumstances (site refusals, absent children, unavailable informed consents, travel arrangements, etc.). They are responsible for the execution of what is the most uncertain, problematic, time-pressured component of this project. Their salaries and percent time basically cover training and data collection that span approximately 12 months, three in budget year 1 and nine in budget year 2. Their salaries and percent time will not vary as a function of the number of sites to be sampled. They will be primarily responsible for hiring the assessors, and it may be quite difficult to get people to do this part-time job.

Training. This phase involves training all data collection personnel in their respective tasks. Three trainers, one for each type of assessment (children’s assessment, classroom ECERS assessment and teacher-child social-emotional-relationship), 24 assessors who will actually conduct the assessments, and six coordinators who will make contact with sites, arrange for informed consent and visits, assign assessors, make travel plans, etc. This training will be held for all relevant project personnel in a single location, one of the Data Collection Headquarter sites, so travel and lodging expenses will only be needed for relevant personnel coming from the other two Headquarter sites. Training will consist of five days, which is really a bare minimum. The Bracken Child Assessment, for example, can be easily taught and practiced in a day, and reliability has been established previously for the instrument and formal reliability assessments probably are not needed because there is very little judgment required of the assessor to score the child’s responses. The ECERS/FDCERS and the teacher-child social-emotional-relationship assessments, however, do require judgments and those require more extensive training and reliability. To be able to do these activities in five days, training videotapes will be used but assessors must practice and demonstrate reliability in real observations before going into the field.

Children’s Pretesting. Each Data Collection Headquarters will have two coordinators and eight assessors, although the actual figure may be different for the three Headquarters depending on sampling. These individuals will be paid \$30 an hour (budget calculations are based on \$240 a day). These individuals may be difficult to hire, so the salary must be

as attractive as possible. They will work in four intensive episodes of one week, five weeks, four weeks, and five weeks spread over a span of approximately nine months. They need to have their own car to travel from site to site and to stay overnight in more distant areas to prevent driving long distances early in the morning. They must deal with personnel at each site, make judgments about child sampling within sites, ensure that informed consent is obtained, conduct the assessments, and be in constant touch with their coordinators. Thus it is approximately a 40% job but one that is worked in full-time segments; if they worked eight-hour days each day, they would earn \$18,000 for a 40%-time position for approximately one year which would be at an annual rate of \$45,000 per year.

Experience in conducting this kind of field work suggests using the following estimates of how many sites and children an assessor could reasonably be expected to conduct in a day. Table 1 indicates 543 sites and assessments to be conducted on 1733 children, a little more than three children per site, but these numbers pertain to the number of *completed* sites and children (i.e., children with both *pre-* and *post-tests* coming from sites having classroom assessments done for each class from which a child is assessed). Therefore, especially at pretesting, the number of sites and especially children to be visited will be greater than the numbers in Table 1 for the completed minimum sample. *Having 24 assessors work for 25 days would mean that each assessor would conduct 5-7 sites per week (one or two a day) and probably test 3-4 children per day*, which gives some room for returning to sites where all children could not be assessed in a single day or other circumstances prevented its completion. While it may seem that an assessor could do more than 1-2 sites per day, this is unrealistic because children take naps and are not available, assessors must return to a site because certain children were absent or informed consent was lacking, the distance between sites can be great especially in rural areas, etc. Children's assessments are expected to take approximately 30 minutes each, but at least 45 minutes should be planned to complete forms, return the child to the classroom and retrieve another, etc. Coordinators will work two additional weeks prior to the assessors going to the field for the purpose of contacting sites, obtaining agreement to participate, sending informed consent materials to the sites, etc. Previous experience (Fiene et al., 2002) indicates that refusal rates are 50%-80% for Head Start, group homes, and family homes but much lower (8%) for child-care centers, so obtaining agreement for sites to participate can be a time-consuming task.

It is difficult to calculate the travel expenses until the actual sample is obtained. Some days, assessors might travel only 20-30 miles to sites in their local vicinity, but on other days they might travel as much as 150-200 miles to sites in outlying areas, so an average of 100 miles per assessment day was used to estimate mileage costs at the government rate of \$0.485 per mile. Also, approximately half of the time the assessors would need to stay overnight to serve sites in outlying areas.

Classroom Quality Assessment. Similar assumptions have been made regarding conducting a classroom assessment. That is, the same number of sites will be involved, a classroom assessment requires approximately one hour of observation plus time to record

the ratings, and centers may have 1-2 classrooms to rate while groups and family-care environments will likely have only one. Thus, *it is assumed that 24 assessors can assess one, occasionally two, sites per day and complete 25 sites in four weeks (20 days)*. Again coordinators must start before the assessors to begin to schedule sites in advance for the assessors to visit.

Post-Test Children's Assessment. The calculations for conducting the post-tests are the same as for the pretests except that the coordinators will work fewer days because sites have already agreed to participate.

Tests. Because the Bracken and ECERS/FDCERS are proprietary tests, the project will need to purchase kits and score sheets for these instruments.

Equipment. Three computers, one for each Data Collection Headquarters, are needed for data entry and to transmit data to the Project Center. In addition, each assessor must be equipped with a cell phone to keep in contact with the coordinators and to verify scheduling with sites.

Supplies. Gifts of children's books or learning materials will be given to sites as a gift for their participation.

Consultants. The project needs to be able to consult with a variety of local and national figures who have experience and expertise in various aspects of this project. These individuals would include specialists who have conducted this kind of field work in Pennsylvania before, individuals who have conducted sampling systems of this sort in Pennsylvania, and data analysis experts. In addition, national consultants, such as Steven Barnett of the National Institute for Early Education Research, Samuel Meisels of the Erikson Institute, and Walter Gilliam at Yale (an expert in state-supported early care and education programs) should be consulted both about the project plans and a draft of the report.

Washington Trip. An amount is listed on this budget for a trip to Washington, which is required only if ACF funding is obtained.

Alternatives to the One-Wave, Minimum Sample Project and Budget

The above approach has two major problems: 1) it is expensive, and 2) it is logistically risky (i.e., can each Data Collection Headquarters indeed hire eight assessors and two coordinators and can this many sites and children be assessed in a 5-week period?). There are two main alternatives to the current one-wave, minimum sample project described above.

Two data collection waves. The logistical risks can be handled by collecting data on the full minimum sample, but over two years (i.e., two data collection waves) rather than one. As discussed in the section on One-vs.-Two-Waves of Data Collection, there are

certain advantages to using two waves in addition to solving the logistics problem. In a Two-Wave design, the total minimum sample could be divided in half, requiring one coordinator and four assessors per Data Collection Headquarters, which would cut the hiring task in half (assuming the same individuals stay for two years). But its full advantages would be achieved if six rather than eight assessors were hired per Data Headquarters site so that somewhat more than half of the total minimum sample would be assessed in each Wave, thus actually increasing the total sample size by 30-50%. As noted above, this would permit more flexibility in sampling so that deficiencies in the first wave in one or another program type could be compensated for in the second wave, and the increased sample size would permit comparisons to be drawn between full-time and part-time children and other program and child care characteristics of interest.

However, this approach costs more, not less, than the One-Wave full minimum sample approach described above. How much more depends on whether the minimum sample is divided in half or actually increased. If the sample is divided in half, data collection expenses are essentially the same but distributed over two years rather than one, but total project personnel would need to work over the entire two-year data collection rather than over only one year of data collection, so total project costs would increase. An approximate estimate of the amount of these costs would increase for a two-year project is to double the budget year 2 amount for total project personnel. **Thus, an estimate of the additional cost of a Two-Wave design is approximately \$168,901 if no additional sites and children are added;** if the sample is increased by 50% (six assessors in each site for two years rather than eight assessors per site for one year, would add \$375,060 to the total data collection costs plus the added Total Project Personnel costs of \$168,901 for a **total increase of \$543,961 for a Two-Wave data collection study with 50% more sites and children.**

The advantages of this approach have been described above, but it also has disadvantages. Obviously it costs more, and an unknown factor is how many assessors and coordinators leave these positions before the two-year project is over. So the Two-Wave approach reduces the logistic pressure and risk and could add more subjects to provide information on more issues, but it costs more, especially in proportion to how many additional children and sites are added.

Eliminate some program types or groups. A strategy that deals with both the cost and the logistical risk is to retain the one wave of data collection but reduce the number of sites and children by eliminating a program type, thus reducing the number of sites and children to be assessed. There are four possibilities: omit the STAR 1 level Keystone STARS sites, the family home care in both Keystone STARS and non-supported child care, the non-supported child care in all three forms, and the family home care in Keystone STARS plus all of the non-supported sites.

- *Omit STAR 1 level.* The rationale for eliminating the STAR 1 level is that these child care centers have really not had much time to benefit from state support. While they are not just entering the system, they are at its lowest level which characterizes most of the units that enter the system. The disadvantages are that eliminating STAR 1

sites would represent a savings of only 54 sites or approximately 10% of the total data collection costs, which do not represent a sufficiently significant savings to warrant this approach.

- *Omit family home care.* This represents a more substantial portion of the total study (31% of the sites), plus visiting family home care sites is more expensive than visiting centers and groups homes because family home involves only six children and there is likely to be only one or two children who meet the preferential criteria for sampling children within a site. Further, they are also likely to be more geographically dispersed, so the actual savings to the project budget is likely to be greater than 31%. Another reason is that the refusal rates in a previous study of classroom quality (Fiene et al., 2002) showed that 64% of family homes and 79% of group homes refused to participate. Refusal rates this high mean that the sites who do participate are likely to be better than average, and thus biasing the results in their favor to the extent that this is true (and such bias is greater for family and group homes than for centers where the refusal rate was only 8%). Disadvantages of this strategy are that family home care represents as large a number of sites in Keystone STARS and non-supported child care, so eliminating family home care means eliminating a category of state-supported care that represents approximately 30% of the different kinds of sites in the Commonwealth (not to be confused with the percentage of children).

The cost saving of eliminating approximately one-third of the sites in the minimum sample is substantial. **Specifically, one would save approximately \$92,855 in training, pretest, test materials, and honoraria the first year and \$134,800 in the second year for total data collection savings of \$227,656.**

- *Omit non-supported child care.* The entire sample of non-supported child care could be omitted, and this would produce a substantial savings of about 34% -- again about one-third of the data collection costs. The rationale for excluding non-supported care is that such sites represent a self-selected group of those who have not volunteered to join the Keystone STARS system, which means they are likely of poorer quality than any of the Keystone STARS sites simply by virtue of this self-selection bias. As a result, they really do not represent a fair comparison and do not represent accurately what quality of care would exist without state support. A more accurate estimate, at least of classroom quality, is provided by the Fiene et al. (2002) study of child care quality in Pennsylvania before Keystone STARS and other supported programs were initiated. While there may be secular trends in the quality of child care that would have operated since 2002 to improve or decrease the quality of care, this comparison is less biased than the non-supported programs. Also, keeping STAR 1 level in the Keystone STARS sample also represents a kind of baseline condition, at least in terms of minimum (but not the lowest) quality of care. The arguments against eliminating the non-supported programs are that using the 2002 data represents a disparity of time and thus a confound and that there are no child assessments to go with the 2002 data leaving STAR level 1 in Keystone STARS as the likely baseline condition.

In general, if one must cut, this seems like a good candidate, because it is a biased group to begin with and because it represents a substantial portion of the cost of data collection (approximately one-third, the same as for the family home sites).

- *Omit both non-supported child care and family home care in Keystone STARS.* The arguments for and against this combined cut are the same as those for omitting family home sites and non-supported sites described above. The cost savings is more substantial, since these two components constitute nearly 50% of the cost of data collection.
- *Summary of Project Costs under Various Alternatives.* Table 8 presents the cost in each of the three budget years for the project under the various alternatives described above. Omitting either the family home sites or the non-supported sites produces a total project savings of only approximately 16% because of the constant costs for overall project personnel. In contrast, omitting both the non-supported and the family home sites produces a total savings of 33%, double that of omitting either of these two program types alone (this occurs because a coordinator in each site can be eliminated under this option but not easily when each category is omitted alone).

Table 8. Reduced Project Comparative Costs for One-Wave Data Collection

	Year 1	Year 2	Year 3	Total
Full Minimum sample	\$501,543\$47 7,196	\$754,130	\$135,113	\$1,390,786
Omit Family Home Sites	\$408,688	\$619,330	\$135,113	\$1,163,131
Omit Non-Supported Sites	\$408,688	\$619,330	\$135,113	\$1,163,131
Omit Non-Supported and KS Family Home Sites	\$310,793	\$482,388	\$135,113	\$928,293

If ACF supported project
omitting Non-Supported and
KS Family Home sites:

Total Cost	\$310,793	\$482,388	\$135,113	\$928,293
Required Cost Share (20%)	\$62,158	\$96,478	\$27,023	\$185,659
ACF Maximum	\$248,635	\$200,000	\$108,090	\$556,725
Actual Cost Share	\$62,158	\$282,388	\$27,023	\$371,569

In the event of ACF support. In the low probability event that an application can be written to ACF by July 7 and this project is one of 3-6 projects in the nation that ACF supports this year, the required cost sharing that ACF would demand is 20% of the total cost of the project (not of ACF's share). Further, ACF caps the total amount they will

provide at \$250,000 the first year and \$200,000 on each of the next two years. Thus, at the bottom of Table 8, we have taken the example of omitting both the non-supported and Keystone STARS family home sites as the illustrative project and then listed its total costs followed by the 20% cost share that ACF would require. Under that, we calculate the amount ACF would contribute in each year as well as the cost share that actually would be required to either meet the 20% minimum or the total cost of the project.

Thus, if no federal support were obtained, the reduced project omitting non-supported as well as Keystone STARS family home sites would cost approximately \$928,293; in the unlikely event that an application could be prepared in time and that it would be funded by ACF, we estimate that ACF could not provide more than \$556,725 towards this total cost and that an additional \$371,569 would have to be obtained from non-federal sources.

References

- Arnett, J. (1989). Caregivers in day care centers: Does training matter. *Journal of Applied Developmental Psychology*, 10, 541-552.
- Caldwell, B. M., & Bradley, R. H. (1984). *Home Observation for Measurement of the Environment*. Little Rock: University of Arkansas at Little Rock.
- Fiene, R., Greenberg, M., Bergsten, M. Fegley, C., Carl, R. & Gibbons, E. (2002). *The Pennsylvania Early Childhood Quality Settings Study*. Harrisburg, PA: Commonwealth of Pennsylvania.
- Harms, T., Clifford, R. M., & Cryer, D. (2005). *Early Childhood Environmental Rating Scale, Revised Edition*. New York: Teachers College Press, Columbia University.
- McCall, R. B. (1979). The development of intellectual functioning in infancy and the prediction of later IQ. In J. D. Osofsky (Ed.), *Handbook of infant development* (pp. 707-741). New York: Wiley.
- McCall, R. B. (1993). Developmental functions for general mental performance. In D. K. Detterman (Ed.), *Current topics in human intelligence, vol. 3* (p. 3-29). Norwood, NJ: Ablex.
- McCall, R. B., Eichorn, D. H., & Hogarty, P. S. (1977). Transitions in early mental development. *Monographs of the Society for Research in Child Development*, 42 (No. 171).
- McCall, R. B., Groark, C. J., & Fish, L. J. (2007). A Caregiver-Child Social-Emotional-Relationship Rating Scale (CCSERRS). Unpublished manuscript, authors, University of Pittsburgh Office of Child Development.
- McCall, R. B., Ryan, C. S., & Green, B. L. (1999). Some non-randomized constructed comparison groups for evaluating early age-related outcomes of intervention programs. *American Journal of Evaluation*, 2(20), 213-226.
- Mehaffie, K.E., & McCall, R. B. (2002). Kindergarten readiness: An overview of issues and Assessment. *Special Report*, University of Pittsburgh Office of Child Development. PA: Pittsburgh.